PRIFYSGOL ABERYSTWYTH UNIVERSITY

Charting Human Microbiome and Metabolome Changes in Disease and Stress

Simon John Cameron B.Sc

Submitted in Full Candidature for the Degree of Doctor of Philosophy

Institute of Biological, Environmental and Rural Sciences

Aberystwyth University

April 2015

IBERS

Athrofa y Gwyddorau Biolegol, Amgylcheddol a Gwledig
Institute of Biological, Environmental and Rural Sciences

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed: _____

Date: _ _ / _ _ / _ _ _ _

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended

Signed: _____

Date: _ _ / _ _ / _ _ _ _

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: _____

Date: _ _ / _ _ / _ _ _ _

E-THESIS DEPOSIT DECLARATION

The content of the electronic copy of this thesis deposited in the electronic repository is identical in content to that deposited in the Library. Where appropriate copyright permission has been sought for the inclusion of third party content.

Signed: _____

Date: _ _ / _ _ / _ _ _ _

ABSTRACT

The role that the human microbiome and metabolome may play in health and disease is well established. To date however, research focus has centred on the human gut microbiome. The microbiome of the respiratory tract has received substantially less attention. Here, the respiratory microbiome of patients with lung cancer and chronic obstructive pulmonary disease, both diseases which have unmet clinical needs, was profiled. For lung cancer, several microbiome-derived biomarkers were detected that may allow, through a non-invasive sampling technique, for the diagnosis of disease status and stage. For chronic obstructive pulmonary disease (COPD), a number of functional differences in the lung microbiome of patients were identified, which may help to explain progression of the disease. Additionally, for lung cancer, metabolomic fingerprinting of sputum from patients revealed a number of metabolites with the ability to differentiate between negative and positive patients, but who presented with clinically-similar symptoms. Using the techniques developed in these clinically-focussed projects, studies of the human microbiome's temporal variability and its response to extreme physiological and environmental stress were completed. Longitudinal sampling of the saliva of healthy individuals over a one year period, showed that the salivary microbiome and metabolome is remarkably stable, but that there is some change in bacterial load. Human saliva is considered a valuable source of both microbiome and metabolome-derived biomarkers, and this work suggests that the comparison of samples collected at different time-points in the year is valid. By sampling members of the Trans-Antarctic Winter Traverse expedition, an analogy to how the human microbiome and metabolome may alter as a result of prolonged human space travel was possible. The salivary microbiome was shown to increase in bacterial load and diversity as a result of the environmental and physiological stresses of the expedition, whilst the stool microbiome was shown to maintain the individual differences evident before the start of the expedition. Additionally, the metabolome of stool, saliva, and blood plasma was shown to be stable throughout the expedition. This research project has shown that analysis of the human microbiome in areas not frequently studied by the field in general can provide novel insights into diseases with unmet clinical needs, and the role that both the microbiome and metabolome can have in the body's response to stress.

ACKNOWLEDGEMENTS

This thesis is dedicated to my wife Rebecca, who has supported and encouraged me throughout the years of my Ph.D work. I could not have done it without her. I would also like to thank my parents, Paul and Jackie Cameron for supporting me throughout my undergraduate studies, and for their unwavering belief and support beyond those years.

I would like to thank my supervisors Professor Luis Mur, Dr Sharon Huws, and Dr Justin Pachebat. Their passion and enthusiasm for research is infectious, and I am grateful for all of their help and support.

Specific colleagues I would like to thank for their help and assistance during the course of this Ph.D are Rob Darby, Kathleen Tailliart, Dr Manfred Beckman, Dr Gordon Allison, Dr Arwel Jones, Dr Hazel Davey, Dr Arwyn Edwards, and Dr Ifat Parveen from the Institute of Biological, Environmental and Rural Sciences (IBERS) at Aberystwyth University, Dr Paul Lewis from Swansea University, Dr Keir Lewis from Hywel Dda University Health Board, and Dr Robert Nash from Phytoquest Ltd.

I am indebted to all of those who agreed to participate in my research studies, including those patients who agreed to donate clinical samples, those who donated saliva at Aberystwyth University, and those who agreed to donate samples in sub-zero temperatures in Antarctica.

I would like to thank those organisation whose funding made this research project possible, namely Aberystwyth University for the provision of an Aberystwyth Postgraduate Research Studentship scholarship and travel fund, Hywel Dda University Health Board and the Society for General Microbiology for funding to attend conferences, and Antarctic Science Ltd for awarding a grant that enabled me to complete the work detailed in Chapter 5 of this thesis.

Finally, I would like to thank my examiners Dr David Whitworth and Professor Eshwar Mahenthiralingam for their time, expertise, and invaluable feedback.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations used throughout this body of work are given here in alphabetical order. Within the body of text the abbreviation is defined at its first appearance.

5-HT	5-Hydroxytryptamine
ANOVA	Analysis of Variance
AUC	Area Under Curve
AVP	Arginine Vasopressin
BAL	Bronchoalveolar Lavage Fluids
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BSTFA	N,O-Bis(trimethylsilyl)trifluoroacetamide
CF	Cystic Fibrosis
CO	Carbon Monoxide
COMET	Consortium for Metabonomic Toxicology
COPD	Chronic Obstructive Pulmonary Disease
CRF	Corticotrophin-Releasing Factor
CT	Computed Tomography
DAMP	Damage Associated Molecular Patterns
ddNTPs	Dideoxy Nucleoside Triphosphates
DES	DNase/Pyrogen-Free Water
DNA	Deoxyribose Nucleic Acid
dsDNA	Double Stranded Deoxyribose Nucleic Acid
DTT	Dithiothreitol
FEV_1	Forced Expiratory Volume in One Second
FTIR	Fourier Transform Infrared Spectroscopy
FT-MS	Fourier Transform Ion Cyclotron Resonance

FVC	Forced Vital Capacity
GABA	Gamma-Aminobutyric Acid
GC-MS	Gas Chromatography Mass Spectrometry
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GWAS	Genome Wide Association Study
HMP	Human Microbiome Project
HPLC	High Performance Liquid Chromotography
IBERS	Institute of Biological, Environmental, and Rural Sciences
IDO	Indeoleamine 2,3-dioxygenase
IFN_γ	Interferon Gamma
IL-6	Interleukin Six
LC-	Lung Cancer Negative
LC+	Lung Cancer Positive
LC-MS	Liquid Chromotography Mass Spectrometry
LDCT	Low Dose Computed Tomography
LLP	Liverpool Lung Project
LTQ-MS	Linear Quadrupole Ion Mass Spectrometry
MG-RAST	Metagenomics Rapid Annotation using Subsystem Technology
MS	Mass Spectrometry
NC	Not Collected
NCBI	National Center for Biotechnology Information
NK	Natural Killer
NMR	Nuclear Magnetic Resonance
NSCLC	Non-Small-Cell Lung Cancer
PAMP	Pathogen Associated Molecular Patterns
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction

QIIME	Quantitative Insights into Microbial Ecology
qPCR	Quantitative Polymerase Chain Reaction
RDP	Ribosomal Database Project
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
ROCCET	ROC Curve Explorer and Tester
rRNA	Ribosomal Ribonucleic Acid
SCLC	Small-Cell Lung Cancer
SNP	Single Nucleotide Polymorphism
TAE	Tris Base, Acetic Acid, and Ethylenediaminetetraacetic Acid
TAWT	Trans-Antarctic Winter Traverse
TNM	Tumour, Node, Metastases
UK	United Kingdom
UKCRN	United Kingdom Clinical Research Network
USA	United States of America
VOC	Volatile Organic Compounds
WHO	World Health Organisation

CHAPTER 1 | Human Microbiomics and Metabolomics in Health and Disease

The sequencing of the human genome in 2001 has allowed a unique insight into the interaction between the human body and its microbial inhabitants (Relman and Falkow, 2001). At least 223 human genes have been identified which have homologues found only in bacteria (Lower, Lower and Kurth, 1996), suggesting a microbial origin through horizontal gene transfer. This suggests the importance of bacteria, viruses, and other microbial organisms, in shaping the genetic evolution of humans and thus, the importance of understanding their genomes and the host-microbiome interactions that exist.

It is widely accepted that the majority of microorganisms are not cultureable under normal laboratory practices; the exact figure predicted to be between 5% to 15%. As Figure 1.1 shows, since the advent of next-generation sequencing technology has allowed for bacterial 16S ribosomal ribonucleic acid (rRNA) sequences to be analysed without the need for prior culturing, the number of predicted bacterial species has increased significantly (Rinke *et al.*, 2013).

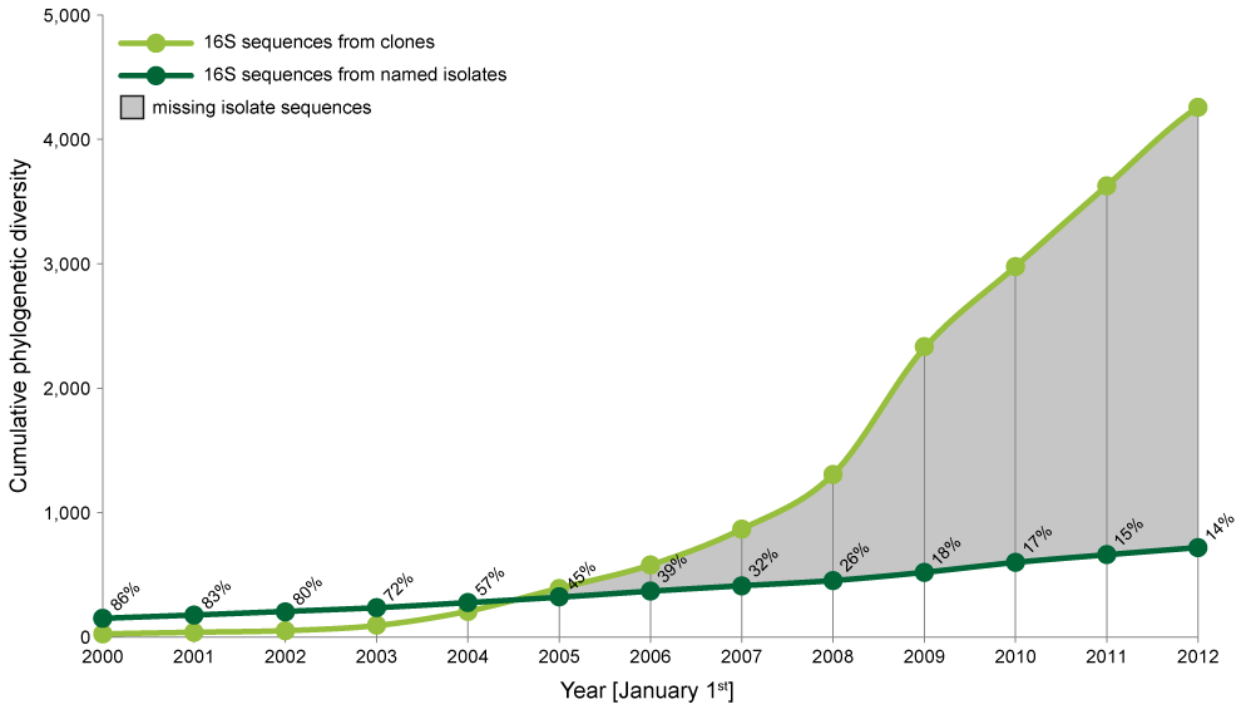


FIGURE 1.1 | Increasing Disparity Between 16S Sequences from Named Isolates and Clones
Next-generation sequencing has enabled the analysis of bacteria without the requirement to culture them beforehand. This has rapidly increased the number of 'predicted' bacterial species (clones) in comparison to those which have been cultured (isolates). Figure taken from Rinke *et al.*, (2013).

The issue of non-cultureable bacteria is paramount in the field of microbiomics as it limits understanding of microbial systems, particularly as unculturable bacteria are not evenly distributed across phyla. This would bias any analysis approach based on culturing. There are multiple reasons for why certain bacteria are considered unculturable, including specific nutrient requirements, pH conditions, atmospheric oxygen levels, or the presence of other microorganisms. Although some progress has been made in improving culture techniques, such as improved natural environment simulation (Vartoukian, Palmer and Wade, 2010), the reliance is still on the use of sequencing techniques to characterise the microbiome in all ecological systems (Schloss and Handelsman, 2005).

These issues constituted the driving force behind the creation of the Human Microbiome Project (HMP). Focusing on metagenomic analysis of microorganisms from across the human body, Figure 1.2, their aim was to build a database of 16S rRNA that characterises a 'normal' human microbiome (Peterson *et al*., 2009). Identifying this has been difficult, and in fact, may prove impossible. Studies into the gut microbiome have found that no single bacterial species was detectable in all gut samples from 154 humans. It may be however, that a core gut microbiome exists in terms of metabolic functional capability rather than taxonomic composition (Kuczynski *et al.*, 2010).

The Human Microbiome Project is developing the resources, both technical and knowledge-based, that will enable microbial communities to be used both as a means of treatment for disease, but also possibly, in its diagnosis. Furthermore, the microbial components of faeces, saliva, sputum and skin swabs have two key features that make it ideal for a diagnostic sample. Firstly, they are easily obtainable, and secondly, they hold a significant resource of molecular information relevant to disease (Sonnenburg and Fischbach, 2011). The purpose of the Human Microbiome Project is to enable the study of variation in the human microbiome and its influence on disease state. One of its most beneficial tasks will be to create a standardised data resource that will act as a reference for researchers exploring changes in microbial populations (Peterson *et al*., 2009).

There are a number of social, ethical and legal implications which must be appreciated and overcome during the Human Microbiome Project. Limited knowledge of the human microbiome may mean that it is hard to fully identify and explain the potential risks and benefits of partaking in studies to participants, which has the potential to cause harm to participants and the reputation of the project. Furthermore, biasing of sampling may lead to a non-representative population of participants; the current Human Microbiome Project only accepts healthy adults between the ages of 18 and 40 as participants. Sampling of participants must also be relatively un-invasive so as to minimise discomfort and biasing of results. These implications are not unique to the Human Microbiome Project, but they must be appreciated for the benefits of the project to be fully realised and further, must also be remembered in studies which

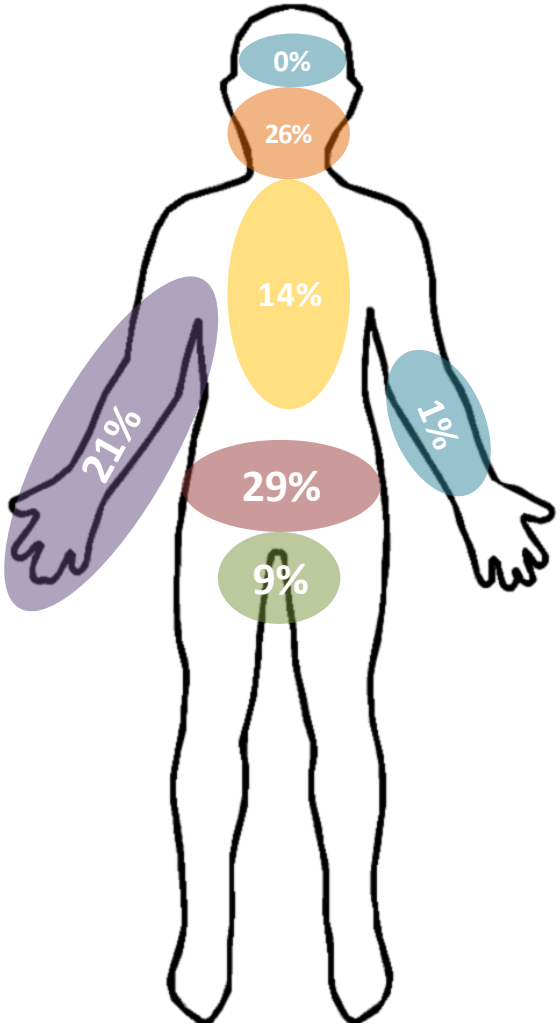


FIGURE 1.2 | Distribution of Bacteria in the Human Microbiome Sequenced by the HMP
Colour coding indicates the percentage composition distribution, by body site, of bacteria that have been sequenced under the Human Microbiome Project, or are currently in the pipeline. These include the eyes (≈ 0%), blood (≈ 1%), gastrointestinal tract (≈ 9%), oral area (≈ 26%), skin (≈ 21%), airways (≈ 14%), and the urogenital tract (≈ 9%). Figure adapted from Peterson *et al.*, (2009).

are seeking to complement it (McGuire *et al.*, 2008).

As the field studying microbiomes has established and developed, so too have the terms used to describe the range of approaches and study features, Table 1.1. As the field moves towards standardisation, to allow for accurate comparisons it is important that consistent terms are used to describe studies into the microbiome. It is likely that as the field expands, and new technologies and approaches are developed, the terms outlined in Table 1.1 will be further developed (Kuczynski *et al.*, 2012).

1.1 | Sequencing and Sampling Techniques in Microbiomics

Since the invention of chain-terminating methods (Sanger, Nicklen and Coulson, 1977), deoxyribose nucleic acid (DNA) sequencing has been at the forefront of nearly all life science disciplines. Many early microbiome sequencing studies employed clone sequencing, whereby the 16S rRNA gene, Figure 1.3, is amplified and then cloned to a high copy number plasmid used to transform *Escherichia coli*. Individual colonies would then be picked and used in single capillary sequencing reactions using chain-terminating,

TABLE 1.1 | Terms in Microbiome Research
In a move towards standardisation in the field studying microbiomes, the terms used to describe the various approaches and technologies have become established. Terms in table taken from Kuczynski *et al.*, (2012).

Term	Description
Amplicon	An amplified fragment of DNA from a region of a marker gene, such as 16S rRNA, that is generated through polymerase chain reaction (PCR)
Metabarcoding	A method for assessing the biodiversity of a system, such as the microbiome, usually through sequencing or other molecular analysis of a marker gene
Metagenomics	The study of the collective genome of microorganisms from an environment
Microbiome	The collection of genes that are harboured by microbiota
Microbiomics	The study of the microbiome
Microbiota	The collection of microbial organisms from a defined environment
Operational Taxonomic Units	Sequences are generally collapsed into these based on sequence similarity, usually set at a threshold of 97%

fluorescently labelled dideoxy nucleoside triphosphates (ddNTPs). Microbiome studies employing Sanger sequencing of clone libraries would usually be in the scale of 1-5,000 reads per study (Gill *et al.*, 2006), and with relatively low numbers of samples due to the sequencing costs involved. Sanger sequencing carried the benefits of long sequence reads, up to 1,000 base pair (bp), and accuracy rates in excess of 99.9% (Shendure and Ji, 2008), allowing identification to the species level of taxonomy. It is however, an expensive method, in both financial and human resources, in comparison to next-generation sequencing technologies and techniques.

The application of next-generation sequencing technologies to the field of microbiomics has allowed for much larger studies, in terms of both sample number and sequence number. One of the main technological advancements was the introduction of the 454 pyrosequencing system (Margulies *et al.*, 2005). This system allowed for the high-throughput sequencing of thousands of 16S rRNA amplicons, from multiple samples in one system run. Sequencing using 454 pyrosequencing suffers from a drawback in terms of reduced sequencing read lengths, usually around 400 bp, compared to traditional Sanger sequencing, and also a high rate of homopolymer errors, making it prone to a higher error rate (Shendure and Ji, 2008). However, the substantial increase in sequence volume, and matched reduction in sequencing cost, saw it quickly replace Sanger sequencing as the primary technology in the field of microbiomics; enabling some of the largest microbiome studies to date to be completed, such as those completed by the Human Microbiome Consortium (The Human Microbiome Consortium, 2012b).

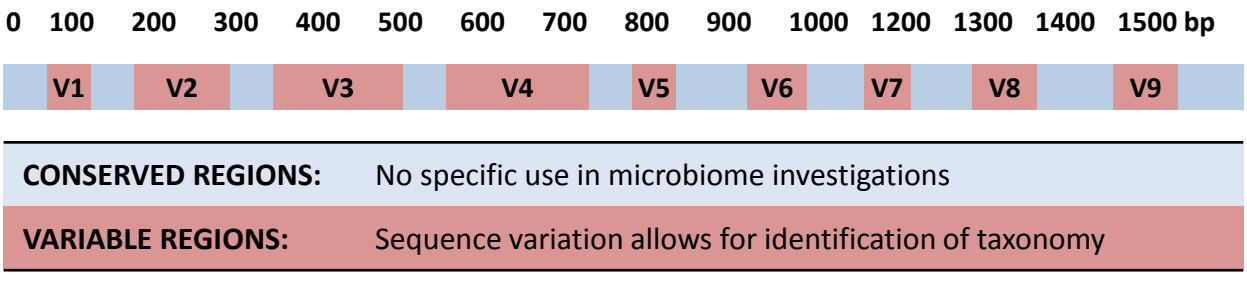


FIGURE 1.3 | The Structure of the 16S rRNA Gene
The 16S rRNA gene is central to current microbiome studies. It is a component of the 30S small subunit of prokaryotic ribosomes and is essential for bacterial cell function. Because of this, it is found in all bacterial genomes, albeit at varying copy numbers. There are nine regions, of various sizes, of the gene that are hyper-variable, regions V1 to V9. Sequencing these regions allows for assessment of phylogeny.

The main limitation of sequencing of the 16S rRNA gene, either through clone sequencing or 454 pyrosequencing, is that information on the microbiome is limited to its taxonomic composition, and usually restricted to the genus level of resolution. The next step in understanding the role that the microbiome plays in health and disease is characterising both its species level composition and functional capacity. This requires a metagenomic approach to sequencing, whereby the entire microbial DNA obtained within a sample is sequenced, rather than just isolated 16S rRNA sequences. As with previous advances in the field of microbiomics, this has been allowed through advances in sequencing technologies. Although 454 pyrosequencing has been used in metagenomic studies, it is somewhat limited by the total number of reads obtainable per run, compared to other available sequencing platforms (Loman *et al.*, 2012). This means that sequencing depth may not be sufficient to identify low copy number sequences, which may have important biological implications.

Metagenomic sequencing techniques are beginning to be applied to the human microbiome, in both health and disease. One of the earliest metagenomic studies catalogued the microbial gene content of the gut of 124 European individuals. A total of 3.3 million non-redundant microbial genes were identified, approximately 150 times as many as contained within the human genome. Within the functional capacity of the microbiome, approximately 40% of genes were common between at least 50% of samples, suggesting a significant element of individual differences exist between microbiomes (Qin *et al.*, 2010).

Microorganisms which do not have a phylogenetic marker gene, as bacteria do in the 16S rRNA gene, such as viruses, can be identified through metagenomic sequencing (Willner *et al.*, 2012b), exploring the wider microbiome. A further step in the application of metagenomic sequencing is the emergence of metagenome wide association studies. For example, in a study of Type 2 diabetes, the taxonomic composition of the gut microbiome varied between individuals, but not between disease and control groups (Qin *et al.*, 2012).

1.1.1 | Bioinformatic Resources for Microbiome Research

The field of microbiomics has advanced significantly with improvements in sequencing technologies. However, advances in this area alone needed to be equalled by advances in sequence analysis methods. With Sanger sequencing, analysis is usually refined to basic local alignment search tool (BLAST) (Altschul *et al.*, 1990) searching of the nucleotide sequence to assign taxonomy. However, the advent of next-generation sequencing has substantially increased the number of DNA sequences in microbiome studies. Improvements in bioinformatic resources needed to match advances in sequencing ability in order to fully utilise the technology in microbiome studies. To this end, a number of bioinformatic pipelines, aimed primarily at high-throughput, barcoded 16S rRNA sequences, were developed.

The Quantitative Insights into Microbial Ecology (QIIME) pipeline has allowed for a standardised approach to 16S rRNA sequence analysis (Caporaso *et al.*, 2010). The move towards metagenomic sequencing of the microbiome has resulted in the development of additional pipelines, backed up by sufficient computing power, to enable an entire metagenomic sequencing project to be taken from raw sequence to functional and taxonomic alignments in one facility (Meyer *et al.*, 2008).

Although analysis pipelines are important in taking raw sequences to biological interpretation, they rely entirely on genomic databases which act as repositories for annotated sequences. For example, 16S rRNA analysis pipelines, such as QIIME, use, amongst others, the Ribosomal Database Project (RDP) database (Cole *et al.*, 2009), Greengenes (DeSantis *et al.*, 2006), or SILVA (Quast *et al.*, 2013) databases to assign taxonomy to sequences. With metagenomic studies, databases which store functional annotations are also required. The Metagenomic Rapid Annotation using Subsystem Technology (MG-RAST) pipeline, for example, relies heavily upon the M5nr non-redundant database (Wilke *et al.*, 2012), which combines protein sequences and annotations from a number of sources.

In addition to the creation and maintenance of annotation databases, an important theme that has emerged within the field of microbiomics is open data sharing of sequences, allowing researchers to

combine datasets from different experiments to further their work. The three main sequence depositories are the European Nucleotide Archive (Leinonen *et al.*, 2011), the DNA Databank of Japan (Kosuge *et al.*, 2014), and the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (Kodama, Shumway and Leinonen, 2012).

1.1.2 | Sampling the Microbiome

Due to the rapid expansion of the field of microbiomics, a diverse range of sampling methodologies have emerged. This diversity has led to difficulties in the comparison of microbiome sequence data from different studies; due to potential bias that may have been introduced at different stages of the analysis pipeline. The common first step, the selection of sampling material, is important to ensure accurate representation of the environment being studied. The oral cavity, for example, offers many sources of sampling, including saliva, tongue, tonsils, teeth, cheeks, and hard and soft palates. Sequencing projects have shown these areas to consist of a defined microbiome, that is not fully shared with other areas (Aas *et al.*, 2005). Therefore, comparing the microbiome of the oral cavity between two studies, one using saliva and one using tongue swabs, would not be a valid comparison of the oral microbiome, but rather a spatial comparison within.

Extraction of genomic DNA is one of the most important steps in microbiome studies. It is however, one of the steps whereby the largest amount of bias can be introduced. For example, if a commercial DNA extraction kit is used, the resulting sequence data is unlikely to be comparable between two different kits, because of differing levels of extraction efficiency (Vishnivetskaya *et al.*, 2014). Differences in sample storage before DNA extraction, however, have been suggested to have minimal impact on resulting sequence data. If any slight difference is introduced by storage practices, these are likely to be surpassed by differences introduced through DNA extraction technique (Wu *et al.*, 2010).

One of the more recent issues to arise in microbiome studies is that of contamination of experimental reagents and kits with microbial DNA. This is a particularly important issue in studies examining

microbiomes with low concentrations of microorganisms, or where the microbiome constituent of interest is at a low concentration (Laurence, Hatzis and Brash, 2014). Additionally, clinical diagnostic techniques using DNA extraction kits could be compromised by contamination with bacteria, such as *Legionella*, during the manufacturing process (van der Zee *et al.*, 2002).

1.2 | Human Microbiomics in Health and Disease

Due to the ease of sample collection, research into the human microbiome originated in the human gut and intestinal system. Following this trend, research in this area is generally more advanced, and crucially, starting to relate bacterial community diversity to physiological conditions, with some explanation of causality.

Sequencing of the core gut microbiome in lean and obese twins has shown that the human gut microbiome has a degree of familial similarity, but that each person has a unique bacterial population composition at the species level of taxonomy. Furthermore, similar levels of co-variation between monozygotic (identical) and dizygotic (non-identical) twins was displayed, though there was a shared core microbiome at the gene level. This study further revealed that a move away from the core microbiome, in terms of phylum-level change and reduced microbial diversity, altered the bacterial genes present and thus resulted in a different physiological state, in terms of obese and lean weights (Turnbaugh *et al*., 2009).

The power of microbiomics to investigate and further understand human diseases was quickly realized after early studies into the gut microbiome, and enabled by rapid advancements in sequencing technologies, and a reduction in their respective costs. Diseases that affect the gut received the majority of focus in terms of the influence of the host-microbiome interaction on disease aetiology, progression, and treatment. Crohn's disease, a chronic idiopathic inflammatory condition of the gastrointestinal tract (Beaugerie *et al.*, 2006), is an early example of this. One of the preliminary studies into the microbiome of Crohn's patients, albeit using only six patients and six healthy individuals and early culture-

independent techniques, suggested that Crohn's disease is characterised by a reduced diversity within the Firmicutes phylum (Manichanh *et al.*, 2006). Later studies utilising advances in sequencing technologies have allowed further insights into Crohn's disease, particularly the phenotypic differences that exist within a disease cohort. Spatial differences within the gastrointestinal tract have been shown, including the disappearance of *Faecalibacterium* and *Roseburia* from the core microbiome in patients with ileal Crohn's disease, and an increase in *Enterobacteriaceae* and *Ruminococcus gnavus* compared to patients with colon Crohn's disease (Willing *et al.*, 2010). This highlights the potential power of microbiomics in improving clinical outcomes as diseases, such as Crohn's disease, could have different treatment regiments developed, such as antibiotic therapy, that target the microbial element of differing phenotypes.

For many diseases, there are evident shifts in the microbiome of affected patients compared to healthy individuals. However, the extent to which the microbiome is a causative factor in disease aetiology and progression, or whether the change is simply a reflection of disease state, is poorly understood. Ever since the link between *Helicobacter pylori* infection and gastric cancer was identified (The EUROGAST Study Group, 1993), the role that the microbiome may play in cancer has received increased attention. For example, colorectal tumour tissue has been shown to form a niche for *Coriobacteria*, whilst showing decreased levels of *Enterobacteria* (Marchesi *et al.*, 2011). However, as the microbiome of colorectal tumour tissue was analysed after carcinogenesis, the causative contribution of the microbiome to the initiation event is not clear. Further work on colorectal cancer has suggested that *Fusobacterium nucleatum* is significantly increased in abundance in patients with colorectal cancer. Again however, sampling of the tumour tissue microbiome after the initiation of tumorogenesis means that a causative link with *F. nucleatum* cannot be established. The pro-inflammatory mechanisms of the bacteria may provide a possible causative factor, but the opportunistic infection of an immune-compromised patient cannot be ruled out (Castellarin *et al.*, 2012).

Approximately 25 diseases and syndromes have been associated with the microbiome of the gastrointestinal tract, albeit at a level of correlation rather than causation. These have included oral and gastrointestinal cancers (Ahn, Chen and Hayes, 2012), colorectal cancer (Sobhani *et al.*, 2011) and Crohn's disease (Manichanh *et al.*, 2006). Nevertheless, increasing numbers of studies are attempting to establish a cause and effect relationship, such as through the application of Koch's postulates to mouse models of inflammatory bowel disease (de Vos and de Vos, 2012). There are multiple mechanisms through which the host microbiome could contribute to a carcinogenic event. For example, the inflammatory response mediated by the host microbiome may lead to localised DNA damage, which may lead to the formation of an oncogene, one of the preliminary stages in carcinogenesis. Additionally, the host microbiome may biochemically alter metabolites, increasing or prolonging the presence of carcinogens. The oral microbiome, for example, has the capacity to convert ethanol to acetaldehyde; a genotoxin capable of causing DNA damage. Normally, ethanol is metabolised to acetaldehyde by alcohol dehydrogenase, and then to acetic acid by aldehyde dehydrogenase. However, the microbiome's contribution has the effect of increasing ethanol to acetaldehyde conversion, resulting in prolonged periods of elevated acetaldehyde levels in the oral cavity and gastrointestinal tract, which may contribute to risk of oral and gastrointestinal cancer risk (Ahn, Chen and Hayes, 2012).

1.3 | Microbiome Changes Associated with Respiratory Diseases

Due to the established link between cystic fibrosis (CF) and the lung microbiome, much of the early work in the field of respiratory microbiomes focused on this disease. However, it was not until early culture-independent were developed that the role of the microbiome in lung disease could be explored. One such method, terminal restriction fragment length polymorphism, was first used to identify a number of bacterial species previously unconnected to cystic fibrosis, a number of which were strict anaerobes (Rogers *et al.*, 2004).

Until relatively recently, it was believed that the human lungs were a sterile environment in terms of microbial communities (Charlson *et al.*, 2011). This was commonly accepted until culture-independent

microbiological techniques demonstrated that the lungs were in fact colonized, in both healthy and diseased individuals, by bacteria (Erb-Downward *et al*., 2011). These findings have opened a new field of microbiome investigation; one which particularly relates to microbial diversity as an indication of diseased state. Currently, the majority of research concerning the lung microbiome has focused on microbial diversity in relation to the lungs and airways of patients with asthma and cystic fibrosis.

1.3.1 | Microbial Diversity in the Asthmatic Lung

Asthma is a respiratory disease which is characterised by a hyper-responsiveness of the airways to various stimuli. This response leads to obstruction of the airways via a combination of bronchial smooth muscle spasm and inflammation; resulting in edema of the respiratory mucosa and mucous secretions. Asthma is usually diagnosed from coughing, wheezing and dyspnea, and commonly treated with steroids inhaled through a nebulizer to reduce airway inflammation (Weinberger and Abu-Hasan, 2007).

Before the wide-spread use of next generation sequencing techniques, studies into the asthmatic lung relied heavily upon the use of bacterial culture techniques. Although limited in its sensitivity to cultureable bacteria, the technique has suggested that there may be a causative link between the lung microbiome and the development of asthma. For example, Bisgaard *et al.,* followed newly born infants for five years, finding that neonates who, at one month old, had their hypo-pharyngeal region colonised with *Streptococcus pneumoniae*, *Haemophilus influenza* or *Moraxella catarrhalis*, are at an increased risk of developing a recurrent wheeze and asthma at an early stage in life (Bisgaard *et al.*, 2007). This study suggests, potentially, that the presence of at least one of these three microorganisms may lead to the development of asthma. However, as this was not uniformly all cases, it may again be that the presence of these microorganisms is an indicator of a predisposed state for asthma, rather than a cause of asthma in itself.

Differences between the microbial populations of a non-diseased lung and an asthmatic lung have been shown, through 16S rRNA amplicon sequencing of 24 adults with asthma (11 adults), chronic obstructive

pulmonary disease (five adults) or no known lung condition (eight adults). A total of 5,054 16S rRNA bacterial sequences, corresponding to a mean of around 2,000 bacterial genomes per cm^2 surface were sampled. The pathogenic Proteobacteria, particularly *Haemophilus* species, were significantly more frequent in the bronchi of adult asthmatics or COPD patients. Furthermore, Bacteroidetes, particularly *Prevotella* species, were significantly more common in control adults than asthmatics or COPD patients (Hilty *et al*., 2010).

The study by Hilty *et al*., was one of the first to identify a difference between the lung microbiome as a whole, rather than individual bacterial species, of an asthmatic and "healthy" individual. However, it was not able to identify whether this difference was a cause or a reflection of a diseased state. This question has been the focus of subsequent studies to evaluate whether a causative link exists. A possible causative link between the lung microbiome and asthma has been suggested by Huang *et al*., who found that when patients with sub-optimally controlled asthma were treated with clarithromycin, they displayed a greater bacterial diversity in their lungs compared to their pre-treatment state. They further found that the patients who responded to clarithromycin had decreased bronchial hyper-responsiveness suggesting that a reduction in bacterial load leads to a lessening of the symptomatic burden of asthma (Huang *et al.*, 2011).

1.3.2 | Microbial Diversity in the Cystic Fibrosis Lung

Cystic fibrosis is an inherited disease which is characterised by the production of thick, sticky mucus as a result of a single defective gene. The defective gene is found on chromosome seven and is a recessive gene, so only homozygous recessive individuals will suffer from the condition. In the lungs, the unusually thick mucus impedes air flow and also provides an ideal growth medium for bacteria and other microbes. As a result, people with cystic fibrosis tend to experience frequent bacterial infections, which are one of the significant causes of mortality (Mahenthiralingam, 2014).

The cause of cystic fibrosis is directly known and unlike as previously discussed with asthma, microbial diversity in the lungs is highly unlikely to be a cause of cystic fibrosis. However, the diversity may instead prove to be a highly valuable diagnostic tool which can be used to aide diagnosis and treatment of bacterial infections. Morbidity and mortality in cystic fibrosis patients are primarily associated with the symptoms of chronic bacterial infections in the bronchial regions of the lungs. These infections were originally thought to be caused by a relatively small number of opportunistic pathogens, such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*. However, recent work using cystic fibrosis sputum samples and next generation sequencing by Guss *et al*., found a significant number of fermenting facultative and obligate anaerobes. These are traditionally difficult bacterial species to identify using culture-dependent techniques, and some are significant opportunistic pathogens. Many of the bacteria identified in cystic fibrosis samples were also present in control samples (Guss *et al.*, 2011), suggesting that the lung microbiome of cystic fibrosis patients is more complex than originally believed; with these under-sampled organisms playing a role that is yet to be elucidated. Understanding how these organisms interact in symptomatic bacterial infections in cystic fibrosis patients will form an important next step in improving treatment options; potentially extending the life expectancy of patients (Mahenthiralingam, 2014).

1.3.3 | Considerations in Lung Microbiome Studies

In undertaking lung microbiomic studies, there are some aspects that should be considered to avoid bias or over-interpretation of the data. The enclosed nature of the lungs presents difficulties in terms of sample collection; and comparisons between different sample materials may be difficult. Thus, some early studies employed bronhcoalveolar lavage fluids as a sampling medium, which later work has suggested samples the lower bronchial mucosal flora, different to that of sputum and bronchial aspirate samples, which represent the upper bronchial tree (Cabrera-Rubio *et al.*, 2012b).

The use of sputum offers a non-invasive method of sampling. Choices are also required into the methods through which the microbiome is defined. Thus, whilst sequencing the hyper-variable regions

of the 16S rRNA gene is widely used for microbiome assessment, due to the limited sequence variation it cannot identify bacteria to a species level. With the advance of sequencing technology and bioinformatics analysis, definition of the entire metagenomic make-up of the lung microbiome in patients with COPD is now possible – thereby providing definitive species-level bacterial identification and insights into the function of the microbiome. These technological advances have already been utilised in the human gut microbiome (Qin *et al.*, 2010), but to date, are yet to be applied to the lung microbiome in COPD patients.

1.4 | Microbiome Changes Associated with Stress

It is well established that exercise-induced stress, that is of a prolonged and/or strenuous nature, can lead to significant perturbations of the human immune system. This is believed to create an 'open window' of increased risk to infection, particularly upper respiratory tract infections. If this stress is combined with other life stresses, such as inadequate nutrition or psychological stress, the overall risk can be substantially higher (Gleeson, 2007). This 'open window' of increased risk to infection suggests that an immune system change reduces resistance to colonisation by opportunistic pathogens. It is likely, therefore, that this reduced resistance also affects microorganisms forming the commensal microbiome in humans; which is present in health.

The human gastrointestinal tract is a system that can be severely affected by stress, both in the short- and long-term. These stressors can be real (physical) or perceived (psychological), and can have an external or internal origin. Although the stress response has an evolutionary advantage, it can have untoward consequences for the gastrointestinal tract. These can include alteration in gastrointestinal motility, changes in gastrointestinal secretions, increased intestinal permeability, reduced regenerative ability of the gastrointestinal mucosa, and negative effects on the gastrointestinal microbiome (Konturek, Brzozowski and Konturek, 2011).

The host response to stress can induce changes in the gastrointestinal microbiome, which can have a knock-on, long-term effect on the host's health. However, the microbiome has been shown be able to modulate the host's immune response, to the benefit of both. In rats, *Lactobacillus paracasei* has been shown to produce metabolites that counteract the stress-induced increase in gut permeability. This could have important implications in the pathogenesis of gastrointestinal diseases, such as irritable bowel syndrome, as probiotics, such as *L. paracasei*, could be used as a prophylactic or treatment regimen (Eutamene *et al.*, 2007).

One of the emerging questions in human microbiomics is the role that the gastrointestinal microbiota plays in the psychological state of the host. Animal studies have suggested that the gastrointestinal microbiome may have a role to play in the regulation of anxiety, mood, cognition, and pain (Cryan and Dinan, 2012). The interaction between the gastrointestinal microbiome and the host brain is mediated through the gut-brain axis; the communication system integrating neural, hormonal, and immunological signalling between the two.

The gut-brain axis provides the microbial constituents of the gastrointestinal tract a potential route through which to indirectly access the brain through biochemical intermediaries. Many gastrointestinal diseases, such as inflammatory bowel disease and irritable bowel syndrome, have a psychological component, meaning the disease symptoms can be worsened by stress, anxiety or depression (Collins, Surette and Bercik, 2012).

The role that stress has on the human microbiome, and the resultant effect that the microbiome has on the host in response, is clearly an important area of research. However, a number of questions still remain outstanding, including the role gender differences may have in the host-microbiome interaction, the gastrointestinal microbiome's effect on physical and psychological development, and whether the microbiome of other human systems also have a role (Foster and McVey Neufeld, 2013).

1.5 | Metabolomics as a Tool for Investigating Disease

In a similar way as the human genome is the collection of all the genes in a human, the human metabolome is the collection of all of the metabolites in the human body. As with microbiome research, any characterisation of the human metabolome is a statistical approximation of the average from a limited pool of individuals. Unavoidably, it may be that individuals have a metabolome that is unique to them, but changes in a person's metabolome may be evident as a result of the body's response to various pressure and changes (Pearson, 2007).

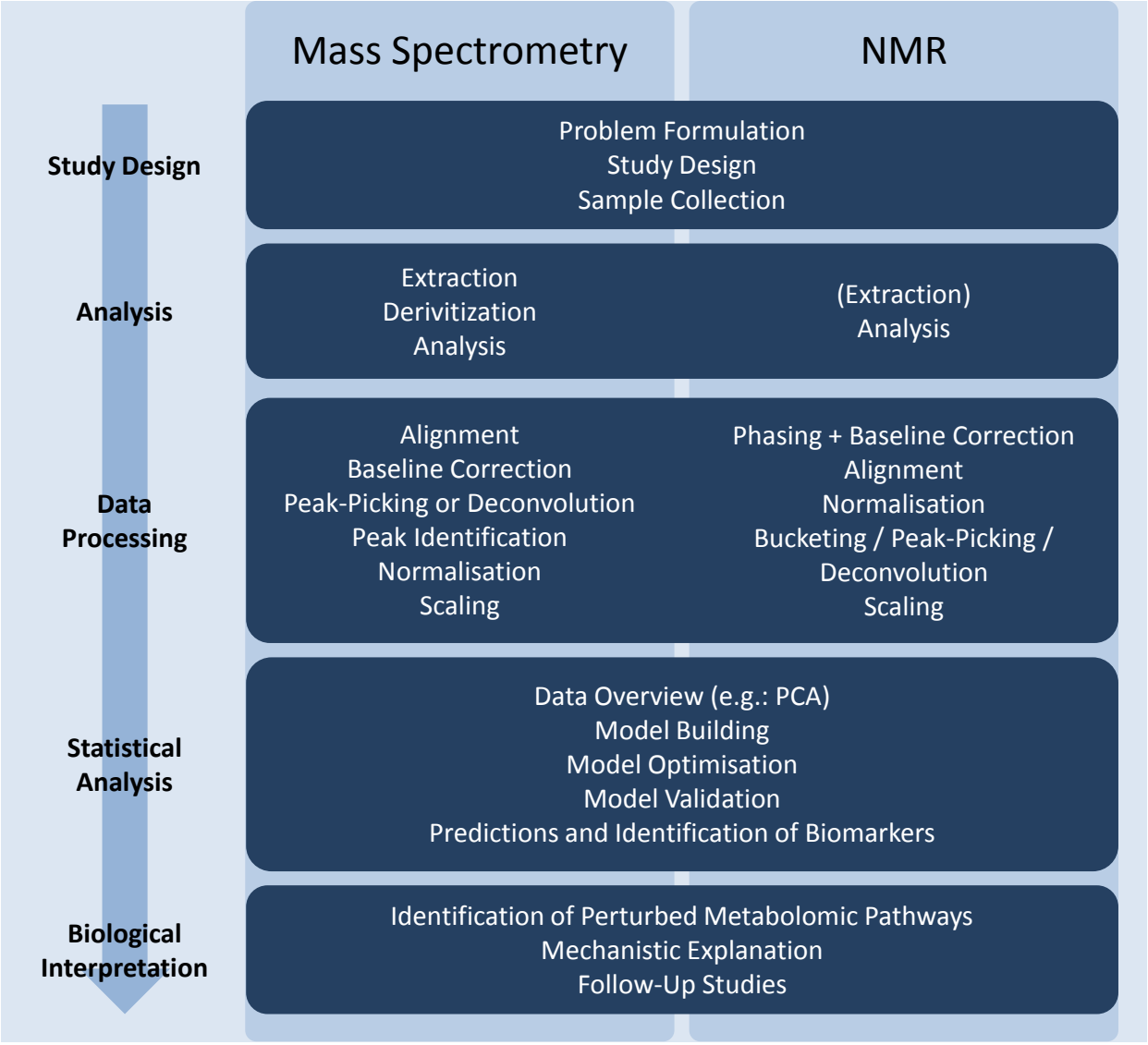


FIGURE 1.4 | Characteristics of Metabolomic Study Pipelines
The majority of metabolomic studies employ a somewhat similar outline. Although the steps can be somewhat similar between different studies, the methods that constitute each step can be markedly different. This will vary depending on the type of technique used, and the starting sample. Figure adapted from Madsen, Lundstedt and Trygg, (2010).

As with the Human Microbiome Project, metabolomic studies are all united by a similar methodology, Figure 1.4, in terms of the steps that are taken from sample preparation to biological interpretation (Madsen, Lundstedt and Trygg, 2010). All metabolomic methods generate large amounts of data, but for studies interested in identifying disease dynamics within the human metabolome, it is likely that only a section of the total metabolome will be altered. For example, mass spectrometry (MS) of saliva samples from 215 individuals, with cancers including pancreatic, breast and oral, identified as little as 57 metabolites which were able to predict disease susceptibility for all three cancers (Sugimoto *et al.*, 2010).

Although significant work has investigated metabolomic changes as a result of cancer, little work has focused particularly on lung cancer. However, respiratory diseases have not been completely neglected, with COPD receiving particular attention in regards to the metabolomic differences evident at different stages of the disease (Ubhi *et al.*, 2012). Metabolomics is a powerful technique in studying the physiological response to, or alterations caused by, disease. It is high-throughput in terms of sample number, with minimal costs in terms of human and financial resources. It does suffer however, from difficulties in comparing data sets generated by differing techniques.

The use of metabolomics in identifying biomarkers for early diagnosis of disease is by no means limited to oncology. In theory, metabolomics could be employed to identify biomarkers for any disease that has the effect of altering the human metabolome. Parkinson's disease, for example, has been suggested as a disease that could be detected through the use of metabolomics, with recent work on patient plasma indicating that 8-hydroxy-2-doxyguanosine could be a biomarker for disease state (Bogdanov *et al.*, 2008). Diabetic kidney disease, which affects around a third of Type 1 diabetes mellitus patients, has also been detected through the use of metabolomics, suggesting the monitoring of Type 1 diabetes mellitus patients through metabolomic analysis of urine could significantly reduce complications associated with the disease (van der Kloet *et al.*, 2012).

1.6 | Methods for Metabolomics

The field of metabolomics began to rapidly expand after the sequencing of the first human genome in 2001, which showed that genetic sequencing does not fully detail disease associated changes. As Figure 1.4 details, there are two main technological approaches employed in metabolomics, nuclear magnetic resonance (NMR) and mass spectrometry. Although different, both can be used in the three main strategies employed in metabolomics, Table 1.2. Each of these have their own individual benefits and drawbacks, and varying degrees of applicability. The 'Gold Standard' is arguable defined as metabolomics, by which as many metabolites within a biological system are identified and quantified, in an unbiased way. This has obvious benefits in that biological pathways can be identified, but it is demanding in terms of resources and technical requirements. The two other approaches either target a small number of previously identified metabolites, or a classification-based approach without identification of metabolites (Madsen, Lundstedt and Trygg, 2010). In the field of metabolomics, there is a move towards a standardised approach to metabolomics, from sample collection to data analysis and reporting, improving the ability to easily compare results from different studies. Although the Metabolomics Standards Initiative has been established since 2007, it is yet to fully create and implement a set of standards required for peer-reviewed publication (Fiehn *et al.*, 2007).

TABLE 1.2 | The Three Main Strategies Used in Metabolomics
The field of metabolomics is not defined by a single strategy, but rather three main techniques, using similar technological methods and sampling, each with its own individual benefits and drawbacks. Table adapted from Madsen, Lundstedt and Trygg, (2010).

	Metabolomics	**Metabolomic Fingerprinting**	**Metabolite Profiling**
Description	Comprehensive analysis with identification and quantification of as many metabolites as possible in a biological system, done in an unbiased way	Fast classification of samples based on metabolite data, without necessarily quantifying or identifying the individual metabolites.	Quantification of a number of pre-defined metabolites
Potential Use	Diagnosis, biomarker discovery, and biological understanding	Diagnosis method	Diagnosis, biomarker discovery, and biological understanding

1.6.1 | Metabolomics Using Mass Spectrometry

Mass spectrometry methods in metabolomics determines metabolites based on the mass to charge ratio in charged particles. The main technologies used in mass spectrometry metabolomics are liquid chromatography mass spectrometry (LC-MS), fourier transform ion cyclotron resonance (FT-MS), and gas chromatography mass spectrometry (GC-MS) (Spratlin, Serkova and Eckhardt, 2009).

Mass spectrometry techniques are generally considered more sensitive than NMR, with LC-MS considered the most sensitive; able to detect compounds at picogram concentrations (Dettmer, Aronov and Hammock, 2007). However, if a particular type of compound is the focus of experimentation, then this might impact upon choice of mass spectrometry technique. Polar molecules, for example, may be detected using electrospray ionisation, though non polar molecules may require a form of pressurised chemical ionisation (Spratlin, Serkova and Eckhardt, 2009).

Volatile organic compounds (VOCs), such as alcohols, ketones, and aldehydes, present issues because they will not easily be extracted using liquid, as in LC-MS, leading to an incomplete profile being analysed. Therefore, GC-MS may be a more appropriate technology because it uses chemical derivitisation, without the need for a liquid extraction method. However, not all metabolites will be derivatised equally, or at all, and therefore GC-MS can be a less sensitive profiling technique that only considers part of the metabolome (Dettmer, Aronov and Hammock, 2007).

It is likely that metabolome profiling, using LC-MS and GC-MS techniques, or derivatives thereof, will only be the first step in 'Gold Standard' metabolomics. These approaches will generate data that can be statistically modelled and individual mass to charge ratios identified that are sufficiently different to, for example, separate disease groups. However, it is unlikely that these methods alone will be sufficient to identify the biological metabolite that is responsible for this difference. A method, such as FT-MS, will be needed to ascertain accurate masses, usually to four to six decimal places, for firm identification (Favé *et al.*, 2011).

The nature of the metabolome, in terms of the significantly different chemistries of metabolites, means that there is no current mass spectrometry platform alone that is able to build an accurate representation of the metabolome from a biological sample. Each mass spectrometry method outlined here has its own distinct advantages and disadvantages. With metabolomes, particularly those from human samples, generally consisting of thousands of metabolites, there is currently no single methodology for analysing them all in an accurate, non-biased way. Therefore, a combination of two or more mass spectrometry techniques is likely to be the best method to give an accurate representation of the metabolome, albeit in separate snapshots (Zhang *et al.*, 2012).

1.6.2 | Metabolomics Using Nuclear Magnetic Resonance

NMR takes advantage of the unique properties of compounds that contain isotopes with an odd number or protons and/or neutrons, and the effect that these properties have on the magnetic spin of the compound. In NMR spectrometry, the most commonly studied nuclei are ^{1}H and ^{13}C, although nuclei from other isotopes have been studied. NMR has a number of advantages and disadvantages unique to the technology. In terms of sensitivity, it is considerably less sensitive than mass spectrometry technologies; being able to detect metabolites at concentrations above or equal to 10 µmol/L. The instrumentation required for NMR is also considerably more expensive than that needed for mass spectrometry (Spratlin, Serkova and Eckhardt, 2009). However, NMR does have a number of advantages in regards to the minimal sample preparation requirement, and the non-discriminatory and non-destructive analysis of the biological sample (Dettmer, Aronov and Hammock, 2007). Arguably one of its main benefits is the relative ease of metabolite identification, which allows for identification of the biological pathway affected by the perturbation, such as disease state, being studied (Griffin, 2003).

1.6.3 | Data Analysis in Metabolomics

As with technological developments in DNA sequencing in the field of microbiomics, similar degrees of advancement in data analysis techniques for metabolomics has matched developments in mass spectrometry and NMR methodologies. Developments of available tools to analyse high quantities of

mass spectrometry or NMR data has arguably been the most beneficial to researchers, allowing unsupervised and supervised approaches to data modelling to be quickly and easily completed (Dettmer, Aronov and Hammock, 2007; Jarvis *et al.*, 2006). Additionally, web based analysis pipelines have been developed that allow for rapid analysis of large datasets, without the need for significant local processing capacity, for both data modelling, such as MetaboAnalyst 2.0 (Xia *et al.*, 2012), or biomarker discovery, such as Receiver Operating Characteristic (ROC) Curve Explorer and Testing (ROCCET) (Xia *et al.*, 2013).

As with microbiomics, reference databases are also essential in allowing identification of metabolites, either through NMR spectra, or mass spectrometry mass to charge ratios or retention times. One of the largest such databases, for metabolites identified in the human metabolome, is the Human Metabolome Database. This contains collections of metabolites with their respective structure, mass, and biofluid location. It also stores the properties of metabolites in regards to which biological pathway they are found in, and in what diseases they have been shown to be altered (Wishart *et al.*, 2013).

1.7 | Human Metabolomics in Health and Disease

In a similar fashion as for the Human Microbiome Project, the Human Metabolome Project was launched as a coordinated effort to build a knowledge base for the field. Starting from a more established foundation of data, as metabolomic research significantly predated the sequencing of the human genome, the Human Metabolome Database, as discussed previously, was launched as a reference point for metabolomic investigation (Wishart *et al.*, 2007).

1.7.1 | Metabolomics in Biomarker Discovery

One of the key principals in biomarker discovery using metabolomics is that disease has an effect on normal bodily function, which translates into a visible difference in the metabolome between those with and without the disease. As Figure 1.5 illustrates, these differences can be visible at many stages of the

pathway from a healthy to a diseased state. For example, biomarkers can be used to detect potential disposition to a disease, biomarkers for the onset of disease, early biomarkers of disease effect, and late biomarkers of disease effect (van der Greef, Stroobant and van der Heijden, 2004). With cancer being a genetic disease, the majority of biomarkers for disease predisposition and risk have a genetic basis. These have traditionally been studied through the use of genome wide association studies (GWAS), and are usually in the form of single nucleotide polymorphisms (SNPs) at specific loci. For example, prostate cancer risk has been shown to be associated with SNPs on chromosomes 3, 6, 7, 10, 11, 19 and X, explaining 16% of the familial risk of the disease (Kote-Jarai *et al.*, 2008), and colorectal cancer predisposition has been linked to a single SNP at chromosome 8q24, which can lead to enhanced Wnt signalling (Tuupanen *et al.*, 2009).

Metabolomics arguably has the greatest potential in identifying biomarkers for the onset of disease, early disease effect, or late effects of the disease. In regards to cancer, for example, this is based upon the 'Warburg Effect'. This principle was developed after early observations of energy source utilisation by cancerous cells, which showed that, even in sufficient levels of oxygen, they use large amounts of

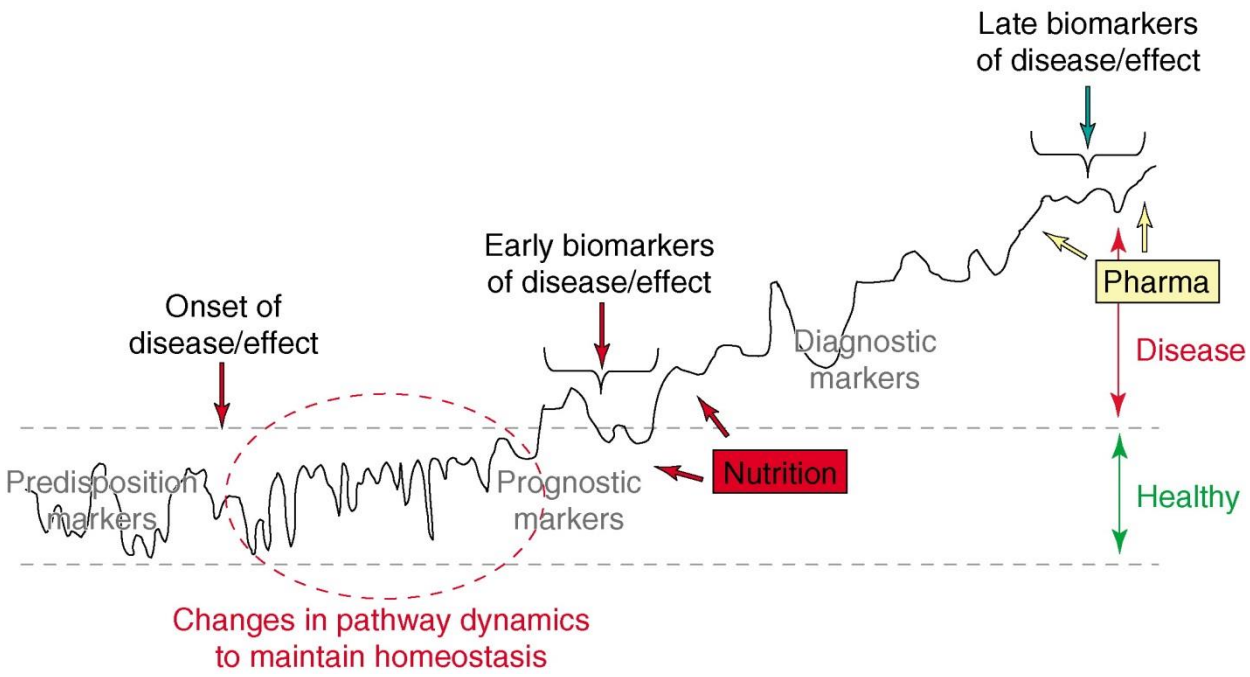


FIGURE 1.5 | The Monitoring of the Pathway from Health to Disease with Biomarkers
Biomarkers can be defined by the stage within the pathway from health to disease at which they are either an indicator of predisposition, indicator of disease onset, or of early disease effect, or of late disease effect which can be used as a target for treatment monitoring. Figure taken from van der Greef, Stroobant and van der Heijden, (2004).

glucose and glutamine in the glycolysis pathway rather than utilising energy through oxidative phosphorylation (Warburg, 1956). This suggests that the presence of cancerous cells, or fully developed tumours, can be detected by the changes that they cause in the level of low molecular weight compounds, which can be detected through metabolomic techniques. For example, patients with bladder cancer have been shown to have significantly different urine, in terms of metabolomic profiles, than those without cancer, which could be used as a screen for early onset of the disease (Issaq *et al.*, 2008). Urine has also been used as a biofluid for identifying patients with kidney cancer, through the use of around 30 potential biomarkers (Kind *et al.*, 2007).

One of the key considerations in metabolomics is the choice of sample to analyse. It is possible that the changes in the metabolome caused by disease presence are localised in the human body. For example, the effects of bladder cancer may be localised to the metabolome of the bladder, and that urine is the best biofluid to use of analysis, rather than blood where metabolomic changes may become diluted to an unobservable level. Additionally, sampling should be as non-invasive as possible to increase the likelihood of the clinical implementation of the method.

1.7.2 | Metabolomics in Drug Discovery and Treatment Monitoring

In oncology specifically, there is a trend towards therapeutic interventions that target particular pathways involved in disease pathogenesis and progression. Biomarkers, derived through analysis of the metabolome, are increasingly being used in the clinical development of novel pharmaceuticals, to identify new molecular targets, to confirm mode of action, and to measure the responsiveness to treatment, drug toxicity, and drug resistance (Spratlin, Serkova and Eckhardt, 2009).

The identification of tyrosine kinase inhibitors (Deininger *et al.*, 1997) as a potential novel therapeutic in oncology has quickly been utilised for the development of multiple clinical therapies. Metabolomics has been used to study the effects of tyrosine kinase inhibitors, to reveal more detail on their potential usefulness. For example, Imatinib, used to treat chronic myeloid leukaemia, has been shown, through

NMR analysis, to prevent the production of essential molecules required for survival of cancerous cells, through the depletion of substrates used in their synthesis (Gottschalk *et al.*, 2004).

NMR has also been used to detect biomarkers indicative of Imatinib resistance. Through NMR analysis, cell lines which show a decrease in mitochondrial glucose oxidation and an increase in phosphocholine levels, later show resistance to Imitinib. If these biomarkers could be monitored in patients' metabolomes, this would have clinical usefulness as it would indicate where a change in therapy is required to ensure continual management of the disease (Spratlin, Serkova and Eckhardt, 2009).

Although the field of oncology has benefited from metabolomics in the identification and validation of novel drug targets, it is by no means the sole example. Microbial infections, such as the Influenza A virus, have also been studied with metabolomics. For example, infection of human fibroblast cells with Influenza A virus, monitored using LC-MS profiling, showed a significant increase in acetylneuraminic acid. This confirms the mode of action of oseltamivir, a leading drug for Influenza A infection, which targets viral neuraminidase (Beyoğlu and Idle, 2013).

An important issue in drug development is that of compound toxicity, as this may impact upon the usefulness of the drug. Indeed, the use of metabolomics to assess drug toxicity has attracted increasing attention. Metabolomics as a field still lags behind both genomics and proteomics in terms of publication number (Robertson, Watkins and Reily, 2011). Metabolomics has been used to monitor drug toxicity and clearance in both animal and human studies. One of the first studies to establish the ability of metabolomics to monitor drug toxicity was performed in rat models of acetaminophen. Using NMR, Clayton *et al.*, (2006) profiled the urine metabolome of rats before treatment with acetaminophen, and found this correlated with the level of liver injury in rats. Additionally, the urine metabolome predicted the degree of drug clearance by the liver through the level of glucoronide levels in the urine (Corona *et al.*, 2012). With the power of metabolomics evident in its ability to predict drug toxicity, the Consortium for Metabonomic Toxicology (COMET) was created to undertake NMR-based metabolomics of urine and

serum to predict the liver and toxicity profiles of novel compounds. The results of resulting COMET studies showed that NMR-based metabolomics can predict organ toxicities at a sensitivity of 41% and specificity of 100% for the liver, and a sensitivity of 67% and sensitivity of 77% for the kidneys (Claudino *et al.*, 2012).

1.8 | Aims and Objectives

The fields of microbiomics and metabolomics offer unique opportunities to study the human body in both health and disease. Moreover, as the fields are both still in their formative years of development, there are still a high number of diseases for which they have not been fully applied. The respiratory diseases lung cancer and chronic obstructive pulmonary disease are two such conditions with an unmet clinical need to develop novel methods of diagnosis, treatment, and monitoring. Investigation of the microbiome and metabolome may help to further understand these diseases. Additionally, techniques from the two fields are rarely combined to understand the host-microbiome interaction in humans.

To this end, the specific objectives of this research project, which each form a subsequent chapter of this thesis, were:

1) Employ microbiome and metabolomic techniques to identify and evaluate novel biomarkers in diagnosing lung cancer.

2) Use metagenomic sequencing to elucidate the structure and function of the microbiome present in the upper respiratory tract of patients with COPD.

3) Combine microbiome profiling and metabolomic fingerprinting techniques to gauge the variability of the human salivary microbiome and metabolome over time.

4) Chart the changes in the human microbiome and metabolome associated with extreme physiological and environmental stress, as an analogy for human space travel.

CHAPTER 2 | Microbiomic and Metabolomic Biomarkers for Lung Cancer

CHAPTER SUMMARY: Developing novel detection methods will improve early diagnosis of lung cancer, increasing the effectiveness of clinical interventions.

Microbiomic Biomarkers for Lung Cancer | In this pilot study, spontaneous sputum samples were collected from ten patients presenting with lung cancer-like symptoms, of which four were diagnosed as positive, and six negative. Nextera® metagenomic libraries were constructed and sequenced on the HiSeq 2500 platform, with subsequent sequences analysed using the MG-RAST pipeline. Taxonomic differences were identified in regards to significant fold changes between negative and positive cases, with five species having significantly higher abundances in positive cases. Functional differences, were evident across a range of biological functions. Regression analyses identified *Granulicatella adiacens*' relationship with six other bacterial species as indicative of positive samples, suggesting it as a potential biomarker. This study offers a novel insight into the functional capacity and species-level taxonomy of the sputum microbiome in patients with lung cancer. Furthermore, it suggests *G. adiacens* as a novel and clinically useful biomarker for lung cancer diagnosis and staging.

Metabolomic Biomarkers for Lung Cancer | The ability of metabolomic fingerprinting to differentiate between patients with and without lung cancer was tested. Spontaneous sputum of 23 lung cancer positive and 11 lung cancer negative patients, alongside 33 healthy controls, was collected. The 67 samples were fingerprinted using linear quadrupole ion mass spectrometry (LTQ-MS) and gas-chromatography mass spectrometry. Analysis based on area under the receiver operating characteristic curve revealed differential metabolites for negative and positive lung cancer, in negative LTQ-MS mode, that had an AUC value of greater than 0.8. This preliminary analysis suggests sputum is a viable sample, and metabolomics has potential as a diagnostic and/or discriminator tool. Furthermore, it can identify specific key metabolites that could aid our understanding of the molecular pathogenesis of lung cancer and possibly guide treatment targets.

2.1 | Introduction

Cancer is not a disease in itself but rather the classification of a group of diseases. These are grouped together because they share common characteristics, such as unregulated cell growth and the spread of cancerous cells from the primary cancer site, to other sites in the body. Due to their varied origins, cancers can have many distinguishing features. Six hallmarks have been identified which are shared by most, and likely all, cancers (Hanahan and Weinberg, 2000). As Figure 2.1 explains, these six hallmarks have been subsequently re-evaluated and two more have been added. Furthermore, two enabling characteristics of cancer, genome instability and mutation and tumour-promoting inflammation, have also been included. Understanding these ten hallmarks of cancer will likely prove fundamental in all areas of cancer research, such as cause, epidemiology, treatment and detection (Hanahan and Weinberg, 2011).

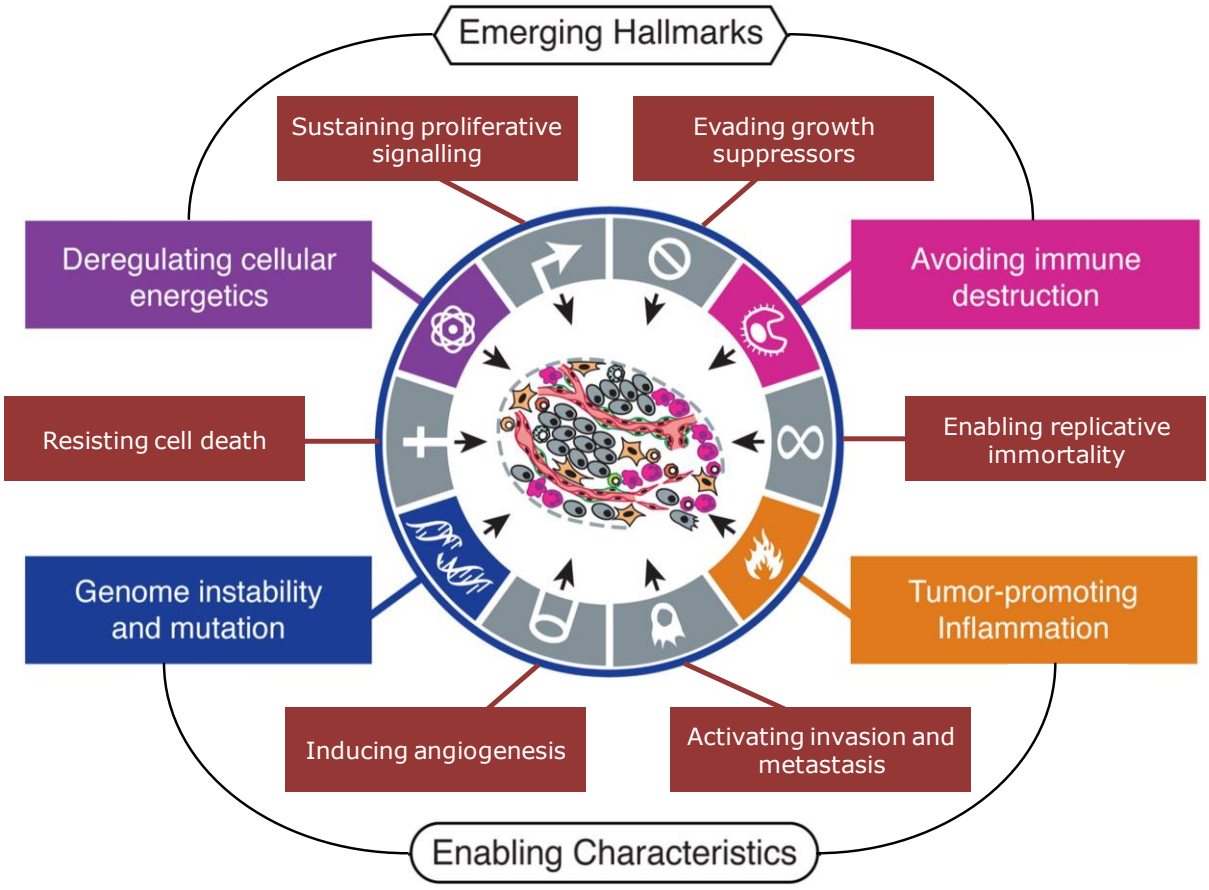


FIGURE 2.1 | Established and Emerging Hallmarks, and Enabling Characteristics of Cancer
The six original (red) hallmarks of cancer, as detailed by Hanahan and Weinberg (2000), in addition to two emerging hallmarks (purple and pink), and the enabling characteristics (blue and orange) as detailed by Hanahan and Weinberg (2011). Figure adapted from Hanahan and Weinberg (2011) to incorporate information from Hanahan and Weinberg (2000).

In 2008, there were approximately 12.7 million new cancer cases, and a total of 7.6 million cancer deaths worldwide (Ferlay *et al.*, 2010). It is commonly accepted that whilst the incidence (new cases) and mortality rates (number of deaths) for the majority of cancers is decreasing in more economically developed countries, they are rising in emerging economies such as China and India. Interestingly, migrant studies have shown that cancer rates in the descendant generations of migrants tend to shift towards the cancer rates of the host country. This suggests that environmental risk factors, such as smoking and weight, rather than genetic differences are responsible for the global variance in cancer statistics (Jemal *et al.*, 2010a).

Lung cancer is a major cause of death in both the developed and developing worlds. It is the leading cause of death attributable to cancer in men, and second only to breast cancer in women. Globally, there are 1.6 million new cases of lung cancer annually, and 1.4 million deaths, as of 2008. This accounts for 12.6% of all cancer incidence and a disproportionate 18.4% of all cancer deaths (Jemal *et al.*, 2010b).

As Figure 2.2 shows, the five year survival rate for lung and bronchus cancer has increased only marginally, in the United States of America (USA), in comparison to all cancers. Furthermore, there is a

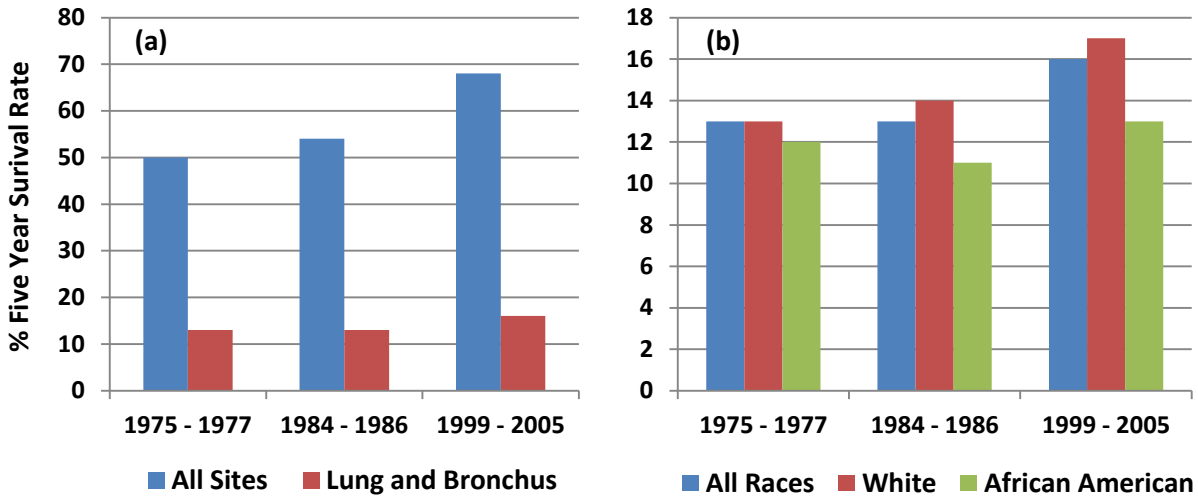


FIGURE 2.2 | Five Year Survival Statistics for Lung and Bronchus Cancer in USA
Figure 2.2a shows the substantially lower five year survival rate for lung cancer, compared to the average for all cancers. The disparity between the survival rate for all cancers, and lung and bronchus cancers is evident, with the latter showing little improvement over the last 30 years, whilst the average for all cancers has improved considerably. Additionally, Figure 2b shows the increasing disparity between the five year survival rate for lung and bronchus cancers for African Americans compared to the average. Figure adapted from Jemal *et al.*, (2010b).

greater divide between the five year survival rates for different groups of Americans, namely 'White' and 'African American' (Jemal *et al*., 2010b). This gap has increased from one percentage point in 1975 to 1977 to four in 1999 to 2005, suggesting the importance of developing the means to identify this cancer at an earlier stage, and thus, increase the prognostic outlook for patients.

Whilst much research into lung cancer has focused on treatment and genetic susceptibility profiling, there is an ever increasing focus on developing the means of identifying patients, with various stages of lung cancer, at an earlier time. Recent work using Fourier transform infrared spectroscopy as a potential high-throughput screen, has suggested that sputum sampling may offer a non-invasive method of screening for lung cancer (Lewis *et al.*, 2010).

2.1.1 | The Cause of Lung Cancer

Lung cancer is one of the few cancers where the major risk factor, in the majority of cases, is clear. In the vast majority of lung cancer cases, 90%, smoking is the primary cause. The risk of lung cancer from smoking increases in a cumulative fashion, with the main indicator of risk being duration rather than quantity of smoking (Cornfield *et al.*, 2009).

The mechanisms of how tobacco smoking can lead to lung cancer were elucidated many years after the initial link between tobacco and lung cancer was made. Only a minority of the overall smoking population develops lung cancer, suggesting that there is a level of individual differences in susceptibility. A study involving 732 clinical patients, half with lung cancer and half without, showed that there was a statistically significant, two-fold difference between the DNA repair capacities of lung cancer patients to the clinical control group. This suggests that carcinogenic chemicals damage DNA in such a way that cancerous cells form and develop into lung cancer (Wei *et al.*, 2000). Oxidative damage has also been shown to cause DNA lesions, such as 8-oxoguanine, which can lead to lung cancer developing because of genetic mutations resulting from mis-match repair (Paz-Elizur *et al.*, 2003). These two studies are examples of work which have added to the substantial body of evidence that shows a causal

relationship between cigarette smoking and lung cancer. The DNA damage caused by carcinogenic chemicals in tobacco smoke lead to mutations in the genetic code, which can cause normal cells to become cancerous, based on the hallmarks of cancer discussed in Figure 2.1.

Radon gas is the second main cause of lung cancer, albeit a comparatively small number, cases annually, and these are mainly concentrated in underground miners, though with some cases of residential radon exposure leading to lung cancer. In total, approximately 18,600 cases of lung cancer annually result from radon exposure in the United States (Field *et al.*, 2000).

2.1.2 | Types of Lung Cancer

Determining the type of lung cancer that has developed in a patient is a key component of determining the correct treatment and management pathway. For lung cancer, two broad classes of classification exist: non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). Patients with SCLC frequently differ from those with NSCLC in terms of their clinical presentation, in that their symptoms are related to the effects of distant metastases (Travis *et al.*, 2004).

Patients with NSCLC are usually classified as one of three main subtypes: adenocarcinoma, squamous-cell carcinoma and large-cell carcinoma. Of these, adenocarcinoma is the most common and is characterised by overproduction of mucin. Squamous-cell carcinoma is the second most common form of lung cancer and typically occurs in the centre of the lungs. Large-cell carcinoma is less common and is characterised by cancerous cells that are large, with excess cytoplasm and large nuclei. The extent of NSCLCs is reported using the Tumour, Node, Metastases (TNM) format, which is important for prognosis and treatment planning. The TNM format ranges from Stage 0 to Stage IV, with the relevant stage determined through assessment of the primary tumour, involvement of regional lymph nodes, and the extent of distant metastasis against set criteria (Travis *et al.*, 2004).

Small-cell lung cancers are less common than NSCLC, with approximately 10% of all lung cancers classified as SCLC. These lung cancers are characterised by their small cells, with minimal cytoplasm, and poorly-defined cell borders. Cancerous cells are usually rounded, oval and spindle-shaped. Typically, patients with SCLC present when the disease has metastasised from the lungs and symptoms frequently reflect this. Small-cell lung cancers are staged differently to NSCLC. Although the TNM format can be used, it does not predict survival and other outcomes well. Typically, SCLC is staged as either limited or extensive disease, with the latter equivalent to Stage IV of the TNM staging format for NSCLC (Travis *et al.*, 2004).

2.1.3 | Diagnosis and Screening of Lung Cancer

It is a significant issue in the diagnosis of lung cancer, that the symptoms of the condition are similar, or even the same, as those for other less serious conditions. This makes the pre-clinical diagnosis of lung cancer particularly difficult because the observed symptoms can be confused with those of other respiratory conditions.

There is currently no recommended test for early detection of lung cancer in asymptomatic patients. However, because there is a clear cause of lung cancer, tobacco smoking in 90% of cases, some high-risk patients may choose to pursue the potential of early diagnosis. However, because of the relatively primitive nature of lung cancer screening techniques, the American Cancer Society stresses the importance of shared decision making between a patient and clinician (Smith *et al.*, 2010). This is of high importance because current methods of early detection for lung cancer can have indeterminate findings, and a high rate of false positives and negatives. This poses the possibility of psychological distress for a patient, leading to increased anxiety around the condition (Byrne, Weissfeld and Roberts, 2008).

Lung cancer has a poor prognosis, with the five-year survival percentage being approximately 15%, Figure 2.2. This has remained relatively static compared with other cancers, and one of the main reasons

for this is that when it is usually diagnosed, the cancer has developed to such a stage where surgical or other clinical interventions usually prove ineffective (Baldwin *et al.*, 2011). The mammography procedure used commonly in breast cancer screening has arguably been a significant driving force behind recent decreases in breast cancer mortality. It has been implemented in many countries because it meets the criteria set out by the World Health Organisation (WHO), initially developed by James Wilson and Gunner Jungner in 1968. These criteria, Table 2.1, are a set of ten principles which should be met by a screening procedure in order for it to be beneficial and cost effective (Andermann *et al.*, 2008).

Currently, possible screening techniques for lung cancer do not meet all of the ten conditions laid out in Table 2.1. The majority of the conflicts with the criteria arise in relation to cost, whom to treat as patients, whether a suitable test or examination exists, and whether the screening technique is acceptable to the population. The latter may be of particular concern as many of the possible screening techniques centre on the use of X-Rays, which may be harmful to patients.

TABLE 2.1 | Wilson and Jungner 1968 Screening Criteria
Ten criteria were set out by James Wilson and Gunner Jungner when they completed their World Health Organisation commissioned report in 1968 on the criteria needed for a national screening programme to be successful. Table adapted from Andermann *et al.*, (2008).

1	The condition sought should be an important health problem
2	There should be an accepted treatment for patients with recognised disease
3	Facilities for diagnosis and treatment should be available
4	There should be a recognised latent or early symptomatic stage
5	There should be a suitable test or examination
6	The test should be acceptable to the population
7	The natural history of the condition should be adequately understood
8	There should be an agreed policy on whom to treat as patients
9	The cost of case-finding should be economical in relation to possible medical costs
10	Case-finding should be a continuing process and not a “once and for all” project

Identifying a target screening population will be of paramount importance in ensuring any lung cancer screen is cost-effective, and remains acceptable to the population at large. Because lung cancer is one of the few cancers which has a definitive cause in a vast majority of cases, identifying those at high-risk of developing the disease is possible.

Work in the Liverpool Lung Project (LLP) has progressed towards developing a risk model for lung cancer patients. By identifying those at the highest risk of developing the disease, the hope is that these people would form the screening population for a lung cancer screen and thereby reduce costs and resources required. By combining data from 579 lung cancer cases with 1157 age- and sex-matched population-based controls, a procedure for creating risk profiles was created. For example, a 77 year old male with an occupational exposure to asbestos, but who is a non-smoker though with a familial history of lung cancer, had an absolute risk of 3.17% for developing lung cancer. By setting a cut-off figure for surveillance of 6.0%, a sensitivity of 0.34 and specificity of 0.90 was achieved (Cassidy *et al.*, 2008). This suggests that this method could easily be employed to identify a high-risk group of patients for whom screening for lung cancer would be beneficial.

Currently, there is a narrow range of large-scale screening procedures which are being trialled in the United Kingdom. These focus primarily on the use of X-Rays, usually in the form of computerised tomography scans. The use of computed tomography (CT) scans has been trialled in many countries, albeit with inconclusive results. For example, a Japanese evaluation of low-dose chest CT (LDCT) scans, using a combined total of 61,914 CT scans, found that of 25,385 patients screened for lung cancer, 210 patients with primary lung cancer were identified. Of these 210, a total of 203 underwent surgery, resulting in a five year survival rate of 90% for all patients and 97% for those on stage IA (Nawa *et al.*, 2012). These results suggest that the majority of lung cancers identified through CT scans were treatable, though the cost of completing CT scans to identify only 0.83% of those with lung cancer may make the cost of using CT scans prohibitive for a national screening programme. However, the benefits of CT scans, such as spiral CT scans, may have been over-estimated. For example, an Italian study

showed how spiral CT scans increased the number of Stage I lung cancer patients identified compared to a randomised control group. However, a three year follow-up suggested that there was a minimal difference between the survival rates of those lung cancer cases identified via spiral CT scans compared to those diagnosed by conventional means (Infante *et al.*, 2009).

The potential benefits posed by the use of CT scans in screening for lung cancer cases at an earlier stage are inconclusive, and may not prove to be clinically useful. This means that different frontiers for screening methodologies for lung cancer need to be explored.

2.1.4 | The Microbiome in Cancer

Since the link between *Helicobacter pylori* and gastric cancer was identified (The EUROGAST Study Group, 1993), the possible links between the host and its microbiome, in terms of response, exacerbation or even the initiation of carcinogenesis are receiving increased attention. Changes in the bacterial loads for key species, for example, have been linked to oral squamous carcinoma, colorectal cancer and oesophageal cancer (Helicobacter and Cancer Collaborative Group, 2001). Within the context of lung cancer, a link between *H. pylori* seropositivity and risk of lung cancer has been investigated through the use of serum samples from patients with lung cancer and age-matched controls (Khan, Shrivastava and Khurshid, 2012). Although, no correlation was reported, it did show that a number of people with lung cancer tested seropositive for *H. pylori* and there is a possibility it could be present in the lung cancer microbiome. The use of serum in this study highlights how the microbiome-cancer links have been investigated using cancers, such as oral (Koshiol *et al.*, 2012; Narikiyo *et al.*, 2004; Sasaki *et al.*, 1998; Tateda *et al.*, 2000) and colorectal (Shiga *et al.*, 2001; Sobhani *et al.*, 2011; Ohigashi *et al.*, 2013) where sampling can be minimally invasive. However, the enclosed nature of the lung complicates sample collection and has involved sampling using bronchoalveolar lavage fluids, tissue from excised lungs obtained during transplantation surgery (Chen *et al.*, 2012), or indirectly through serum (Koshiol *et al.*, 2012).

Sputum is a complex of mucus, pathogens, cellular debris and other particles trapped in the lungs by mucus. It provides a non-invasive method of obtaining upper respiratory tract samples that also involves minimal patient discomfort (Lewis *et al.*, 2010). As sputum is symptomatic of inflammatory lung airway diseases such as COPD, asthma, chronic bronchitis, and cystic fibrosis, it is often used to provide insight into the underlying pathogenesis (Voynow and Rubin, 2009). Indeed, conditions such as asthma (Hilty *et al.*, 2010), COPD (Erb-Downward *et al.*, 2011; Pragman *et al.*, 2012; Sze *et al.*, 2012) and cystic fibrosis (Willner *et al.*, 2011) have used microbial profiling techniques to reveal potentially important insights into the role that microbes may play in disease aetiology, progression and treatment.

Sputum from lung patients has been used to explore, albeit in a culture-dependent method, the microbial flora and the level of antibiotic resistance (Li *et al.*, 2013). A further, culture-independent study using amplicon sequencing suggested that, in sputum, there are significant differences between lung cancer patients and controls, particularly within the *Granulicatella*, *Abiotrophia*, and *Streptococcus* genera (Hosgood *et al.*, 2014). However, to date, there has been study into the metagenomic composition of the sputum microbiome in lung cancer. Therefore, resolution at the species level of taxonomy has not been possible, and the functional capacity of the microbiome has not been investigated.

2.1.5 | Metabolomic Insights into Lung Cancer

An alternative screening methodology to radiography is the utilisation of molecular markers, both genetic and metabolomic, in biofluids. For example, microRNAs have been widely studied and suggested as potential biomarkers for lung cancer in sputum (Xie *et al.*, 2010), plasma (Shen *et al.*, 2011), and serum (Foss *et al.*, 2011) samples. In fact, for a number of reasons microRNAs are widely considered to be one of the most promising areas for lung cancer biomarkers currently in development. For example, microRNAs have been shown to be highly stable in serum and plasma samples, the method for obtaining a sample is minimally invasive, and the analysis of microRNAs using quantitative polymerase chain reaction (qPCR) is well established. However, a number of questions still remain regarding microRNAs as

biomarkers for lung cancer. One of the most pressing is the source of microRNAs, as their cellular origins are still unknown, and it is still ambiguous whether they are produced by cancerous cells or by healthy cells in response to the presence of cancerous cells (Wittmann and Jäck, 2010).

The ease of analysis of biofluids using mass spectrometry or NMR makes metabolomics a well suited methodology for identifying non-invasive biomarkers of lung cancer. The focus, to date, of metabolomics in lung cancer has been on the evaluation of serum, urine, or tumour biopsies. Analysis of serum, for example, through both LC-MS and GC-MS approaches, has suggested it as a source of biomarkers for lung cancer. A small-scale pilot study sampling lung cancer patients before and after surgical intervention, alongside patients without lung cancer has suggested approximately ten candidate biomarkers for lung cancer, including sphingosine, oleic acid, and serine (Chen *et al.*, 2014).

Urine is arguably one of the least invasive biofluids to collect, and has a proven track record as a source of biomarkers for a number of diseases. A large scale study of approximately 500 lung cancer patients and 500 population controls employing LC-MS identified creatine riboside and *N*-acetylneuraminic acid as potential biomarkers for early stage lung cancer. Additionally, analysis of the cancerous tissue metabolome and the metabolome of nearby non-cancerous tissue confirmed that both metabolites were significantly enriched in cancerous tissue; revealing the direct association of the biomarkers with lung cancer (Mathé *et al.*, 2014). However, urine has a comparatively low level of metabolite data compared to other biofluids. Therefore, potential biomarkers that are present at a low level in the body, may be at an undetectably low level in the urine.

One of the least studied biofluids in lung cancer metabolomics is sputum. Sputum is a complex of mucus, bronchial cells, and microbial cells that is commonly produced by those with various lung conditions, including lung cancer and COPD. Cytological analysis of sputum has long been a diagnostic tool in lung cancer, that is commonly paired with radiography, and has a sensitivity of around 60% and specificity of >99% (Erkiliç, Özsaraç and Küllü, 2003). The DNA methylation and microRNA profiles of

sputum has been suggested as a potential biomarker in lung cancer but these are by no means established as reliable biomarkers (Hubers *et al.*, 2013; Guzmán *et al.*, 2012). The microbiome of lung cancer patients has not been investigated fully, particularly in regards to the species level of taxonomy and functional capacity, which can only be determined through metagenomic techniques.

The fingerprinting approach of fourier transform infrared spectroscopy (FTIR), as a non-invasive technology to detect lung cancer in sputum samples, has been carried out. Through this, FTIR was shown to have the potential to act as a non-invasive, high-throughput and cost-effective for screening sputum samples from patients at high risk of developing lung cancer. Additionally, it supported the role that sputum could play as a biofluid for use in lung cancer screening (Lewis *et al.*, 2010).

2.1.6 | Aims and Objectives of Chapter

The microbiome and metabolome of sputum from patients with suspected or confirmed lung cancer have received minimal attention to date. The aims and objectives of this portion of work were to:

1) Employ metagenomic sequencing to characterise the species-level taxonomy, and functional capability of the lung cancer microbiome represented by sputum.

2) Explore the lung cancer microbiome as a potential source of novel biomarkers for early-stage identification of lung cancer, and possible later clinical staging of the disease.

3) Characterise the metabolome of patients with suspected lung cancer, using a metabolomic fingerprinting approach, to identify novel biomarkers that may be able to differentiate a group of clinical patients with suspected lung cancer into those negative and positive for the disease.

2.2 | Materials and Methods

The Medlung study received ethical approval from the loco-regional ethical committee (05/WMW01/75). Informed consent was obtained from all patients prior to giving a sputum sample. All samples and data were link anonymised prior to processing and subsequent data analysis.

2.2.1 | Patient Recruitment and Sampling

Spontaneous sputum was collected from ten clinical patients, who were referred for further diagnostics at Prince Phillip Hospital, Llanelli, United Kingdom (UK), after presentation with lung cancer-like symptoms at their General Practice, as part of the Medlung observational study (United Kingdom Clinical Research Network (UKCRN) ID 4682). Spontaneous sputum samples were taken before bronchoscopic investigation for lung cancer diagnosis. All spontaneous sputum samples were confirmed as sputum, based on bronchial cell content, by a Consultant Pathologist in the Hywel Dda University Health Board Pathology Service.

2.2.2 | Processing of Raw Sputum

After collection, raw sputum samples were frozen at -80°C for up to seven days, at which time they were defrosted in ice for approximately one hour. Sputum cells were isolated as described by Lewis *et al.*, (2010) through the addition of 0.5 mL of a working solution of dithiothreitol (DTT), made up by adding 2.5 g of DTT to 31 mL of 30% aqueous methanol, and 5 mL of 30% aqueous methanol, following which they were placed on a vortex mixer for 15 minutes. Samples then underwent centrifugation at 1800 x g for 10 minutes. The supernatant was then removed and the pellet transferred to a clean 1.5 mL microcentrifuge tube and frozen at -80°C.

After processing, those samples undergoing metabolomic fingerprinting were stored at this temperature for up to two years, whilst those undergoing total genomic DNA extraction were stored for no more than seven days.

2.2.3 | Isolation of Total Genomic DNA

Total genomic DNA was extracted from 100 µL of treated sputum using a FastDNA SPIN kit for soil (MP Biomedical, Santa Ana, USA) following manufacturer's instructions. Bead beating was carried out in a FastPrep-24 machine (MP Biomedical) with three cycles at speed setting 6.0 for seconds, with cooling on ice for 60 seconds between cycles. Genomic DNA was eluted in to 30 µL of DNase/Pyrogen-free water (DES) and double stranded deoxyribose nucleic acid (dsDNA) concentration determined using the Quant-iT dsDNA High Sensitivity assay kit and a Qubit fluorometer (Life Technologies, Paisley, UK).

2.2.4 | Metagenomic Library Preparation and Sequencing

After extraction of genomic DNA, samples were normalised to 10 ng/µL with PCR grade water (Roche Diagnostics Limited, West Sussex, UK) and 50 ng used to create metagenomic libraries using the Nextera® DNA kit (Invitrogen, San Diego, USA) following manufacturer's instructions, except that a MinElute PCR purification kit (Qiagen, Ltd Crawley, UK) was used for the clean-up of tagmented DNA. Nextera® DNA libraries were quantified using the Quant-iT dsDNA High Sensitivity assay kit, and approximate library sizes determined by running on a 2 % agarose gel alongside HyperLadder IV (Bioline, London, UK). Sample libraries were pooled in equimolar concentrations and sequenced at 2 x 151 bp using an Illumina HiSeq 2500 rapid run, with samples duplicated over two lanes, and following standard manufacturer's instructions at the IBERS Aberystwyth Translational Genomics Facility.

2.2.5 | Metagenomic Sequence Analysis

After sequencing, output files for each lane were combined into one file, using the BioLinux 7 environment (Field *et al.*, 2006), for each read direction. Sequencing files were uploaded to MG-RAST (v3.2) (Meyer *et al.*, 2008) as FASTQ files. Paired-end reads were joined using the facility available within MG-RAST, with non-overlapping reads retained. Sequences were dereplicated and dynamically trimmed using the default parameters for FASTQ files, and human sequences removed by screening against the *Homo sapiens* (v36) genome, available via NCBI. The MG-RAST pipeline used an automated BLASTX

annotation of metagenomic sequencing reads against the SEED non-redundant database (Overbeek *et al.*, 2005). SEED matches can be matched to identity at various taxonomic levels; including genus and species levels. Organism abundances were modelled and exported from MG-RAST using the 'Best Hit Classification' after alignment to the M5NR database, with alignment cut-off parameters set at an e-value maximum of 1×10^{-5}, a minimum identity of 97%, and a minimum alignment of 15. Functional abundances were modelled and exported from MG-RAST using 'Hierarchical Classification'. SEED matches can also be related to metabolic information, again at different levels of classification. The coarsest level of organization; the generalized cellular function was termed level 1, and the finest, individual subsystems level 3. To normalise for potential variations in sequencing efficacy, sequence abundances were transformed into percentages based upon the total read abundance for each sample. Statistical analysis was completed using the MetaboAnalyst 2.0 (Xia *et al.*, 2012) facility and MINITAB 14 package. Eukaryotic taxonomic classifications were trimmed based on literature searches to remove poorly classified reads. Sequence files can be viewed on MG-RAST via the IDs listed in Chapter 2 Appendix, Supplementary Table 2.1 and raw sequence files have been deposited at the European Nucleotide Archive under primary accession number PRJEB9033 and secondary accession number ERP010087. Unless otherwise stated, all *P* values indicate significance of one-way analysis of variance (ANOVA) tests.

2.2.6 | 16S rRNA Quantitative PCR

Quantitative PCR was carried out on neat extracted DNA against standards created by amplifying the 16S rRNA gene of two randomly selected lung cancer positive and two lung cancer negative samples. This was completed through amplification of the 16S rRNA gene in a 20 µL reaction volume consisting of 10 µL of 2 x BioMix (BioLine), 0.25 µL each of 27f (5'-AGA GTT TGA TCC TGG CTC AG-3') and 1389r (5'-ACG GGC GGT GTG TAC AAG-3') primers (Hongoh, Ohkuma and Kudo, 2003) to give a final concentration of 500 nM, 1 µL of neat extracted DNA, and 9.5 µL of PCR Grade Water (Roche). The reaction volumes were then subjected to PCR consisting of 94°C for two minutes, 30 cycles of 94°C for 45 seconds, 55°C for 45 seconds, and 72°C for 90 seconds, followed by a final elongation step of 72°C for seven minutes.

The resulting PCR products were combined and purified using an Isolate II PCR and Gel Extraction purification kit (BioLine), following manufacturer's instructions, and quantified using an Epoch spectrometer.

The resulting DNA concentration was used to estimate the total number of 16S rRNA gene copies and serial dilutions of 10^{10}, 10^{8}, 10^{6}, 10^{4}, 10^{2}, and 10^{0} made. Quantitative PCR was completed on neat extracted DNA against standards with each reaction completed in 25 µL volumes, each consisting of 12.5 µL 2 x SYBR Green Mastermix (Life Technologies), 0.25 µL of each EubF1 (5'-GTG STG CAY GGY TGT CGT CA-3') and EubR1 (5'-ACG TCR TCC MCA CCT TCC TC-3') primer (Maeda *et al.*, 2003), in a final concentration of 400 nM, 11 µL of PCR Grade Water (Roche) and 1 µL of neat DNA extract. Reactions were run using a C100 thermal cycler (BioRad, Hercules, USA) and CFX96 optical detector (BioRad), with data captured using CFX Manager software (BioRad), under conditions of 95°C for ten minutes, 40 cycles of 95°C for 15 seconds and 60°C for 60 seconds, followed by a melt curve consisting of a temperature gradient of 60°C to 95°C in 0.5°C increments, each for five seconds.

2.2.7 | Linear Quadrupole Ion Trap Mass Spectrometry (LTQ-MS)

After processing, 20 µL of sputum was added to 20 µL of ultrapure water and 40 µL of ice-cold high performance liquid chromatography (HPLC) grade acetone. Samples were placed on a vortex mixer for five seconds, cooled on ice for 30 minutes, and then underwent centrifugation at 11 000 x g for five minutes. A total of 50 µL of the supernatant was removed and transferred to a glass vial, to which 250 µL of 70% methanol (made up using HPLC grade methanol and ultrapure water) was added. All vials were capped and randomised before injection using an autosampler, with tray temperature kept constant at 15°C. For each sample, 20 µL was injected into a flow volume of 60 µL per minute water-methanol, in a ratio of 70% water and 30% methanol, using a Surveyor liquid chromatography system (Thermo Scientific, MA, USA). Data acquisition for each individual sample was conducting, in alternating positive and negative ionisation mode, over four scan ranges (15-110 m/z, 100-220 m/z, 210-510 m/z,

500-1200 m/z) on an LTQ linear ion trap (Thermo Electron Corporation, CA, USA), with an acquisition time of five minutes.

2.2.8 | Gas Chromatography Mass Spectrometry

As with LTQ-MS, 20 µL of processed sputum was added to 20 µL of ultrapure water and 40 µL of ice-cold HPLC grade acetone. Samples were placed on a vortex mixer for five seconds, cooled on ice for 30 minutes, and then underwent centrifugation at 11 000 x g for five minutes, after which 50 µL of the supernatant was removed and dried in a DNA SpeedVac at 40°C until all liquid had been removed. After drying, 30 µL of a 20 mg/ml solution of methoxyamine in pyridine was added to the sample in a glass vial. Vials were capped and incubated at 90°C for 15 minutes. After incubation, 20 µL of N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) was added to the sample, alongside 5 µL of an alkane mix, warmed to 60°C for addition, consisting of C_{10}, C_{13}, C_{15}, C_{18}, C_{19}, C_{23}, C_{28}, C_{32} and C_{36} alkanes, each at a concentration of 2 µL/mL in pyridine, for alkanes which are liquid at room temperature, and 2 mg/mL in pyridine, for alkanes which are solid at room temperature, capped and incubated at 90°C for 15 minutes. After undergoing derivitisation, samples were run on an Agilent 6890N GC (Agilent Technologies, CA, USA) linked to a 5973N MS (Agilent Technologies) with a DB5 equivalent column (Agilent Technologies), using a helium carrier gas at a flow rate of 1 mL per minute, set at a constant rate. GC-MS heating was set at 80°C for three minutes, rising to 280°C at a rate of 15°C per minute, and then to 330°C at a rate of 50°C per minute. The inlet temperature was set at 280°C with a 2:1 split ratio and the MS transfer line was set at 330°C, with an isooctane solvent being used to wash the auto-sampler. Samples were run in duplicate, starting with raw sputum each time and analysed on separate GC-MS runs.

2.2.9 | Metabolomic Data Analysis

To normalise the LTQ-MS data, it was exported into Microsoft© Excel™ 2010 and the total ion count for each sample was used to transform the intensity value for each metabolite in to a percentage of the

total ion count, after the removal of metabolites less than 50 m/z in LTQ-MS profiles. Multivariate analyses were then completed using the PyChem (Version 3.0.5g Beta) package (Jarvis *et al.*, 2006), the MetaboAnalyst 2.0 online platform (Xia *et al.*, 2012) and the ROC Curve Explorer and Tester (ROCCET) online platform (Xia *et al.*, 2013). GC-MS data was exported and the mean values for each data point calculated from duplicate sample runs. Retention times of alkane standards were then removed and data normalised. Data was then modelled as for LTQ-MS data as previously described.

2.3 | Results

Two separate patient cohorts donated samples for this portion of work. A total of ten patients donated sputum samples that were analysed through metagenomic sequencing, and 34 clinical patients and 33 'healthy' participants donated samples which were analysed through metabolomic fingerprinting.

2.3.1 | Participant Cohort for Lung Microbiome Study

After histological investigation of the ten patients referred with lung cancer-like symptoms, four patients were diagnosed with lung cancer (one squamous cell NSCLC, one adenocarcinoma NSCLC, one large cell carcinoma NSCLC, and one where a bronchoscopy was not possible and a radiological diagnosis

TABLE 2.2 | Average Patient Characteristics for Both Negative and Positive Lung Cancer Groups
Data are means for each either the negative or the positive lung cancer groups. Standard deviations are given in brackets. FEV_1 % of predicted is forced expulsion volume of lungs in one second, as a percentage of the predicted value for that patient. Smoking pack years is individual history normalised to 20 cigarettes a day for one year. CO level is carbon monoxide in parts per million concentration.

	Negative (LC-)	Positive (LC+)
Number	6	4
Age	58.8 (14.8)	73.3 (7.2)
Gender		
Male	4	2
Female	2	2
Smoking Status		
Current	4	3
Ex	2	0
Never	0	1
Smoking Pack Years	57.7 (35.4)	28.8 (24.6)
Infection Present		
Yes	0	0
No	6	4
Antibiotic Use		
Yes	1	0
No	5	4
CO Level (ppm)	21.0 (24.5)	7.3 (8.0)
FEV1 % of Predicted	77.1 (22.3)	69.3 (16.0)

was required), and six were found to be negative for lung cancer presence after a one year follow-up period. Summarised patient information is shown in Table 2.2, and full individual patient information in Chapter 2 Appendix, Supplementary Table 2.1. No discernible differences between the two patient groups were observed.

2.3.2 | Preliminary Read Analysis and Bacterial Load

Sequencing statistics, both pre- and post-quality control process are summarised in Chapter 2 Appendix, Supplementary Table 2.2, alongside corresponding one-way ANOVA *P* values. In all but one of the sequencing statistics, identified rRNA features, no significant differences were detected. This suggests that sequencing with the the HiSeq 2500 platform, along with subsequent bioinformatic analysis using MG-RAST did not introduce any bias; which may affect interpretation of results. All ten metagenomes sequenced are publicly available via the links detailed in Chapter 2 Appendix, Supplementary Table 2.1.

In regards to estimated bacterial load, no significant difference (*P* value = 0.616) was observed between the Log_{10} values of estimated 16S rRNA copy number of LC+ (5.63) and LC- (6.06) samples.

2.3.3 | Comparison of Taxonomic Composition of Microbiome

At the taxonomic level, Figure 2.3a, principal component analysis (PCA), created using the MG-RAST analysis platform, appears to show some separation, mainly along principal component one, between the positive and negative lung cancer groups. Additionally, there appears to be no clustering in regards to smoking status, suggesting that lung cancer status may be one of the primary factors in the observable separation based on taxonomic composition of the microbiome.

The ‘core’ microbiome of both negative and positive lung cancer patients, to the species level of taxonomic classification was investigated, Table 2.3, as this is likely to give greater insight into the microbiome than genus. A total of seven species were found to be present in all ten samples, with

Streptococcus viridans found to be significantly (*P* = 0.042) higher in LC+ samples. Six further species were found to be present in all LC- samples, but not all of the LC+ samples, but none were significantly different in their level of abundance. Furthermore, a total of 16 bacterial species were found in all of the LC+ samples but not all LC- samples, with *Granulicatella adiacens* (*P* = 0.015), *Streptococcus intermedius* (*P* = 0.023), and *Mycobacterium tuberculosis* (*P* = 0.036) found to be significantly higher in the LC+ group.

Additionally, at the taxonomic level, significant (*t*-Test *P* value < 0.05) fold changes in regards to species abundance between positive and negative lung cancer cases were identified through use of the MetaboAnalyst 2.0 platform, Figure 2.4. This reflects differences in the 'core' microbiome changes shown in Table 2.3, namely significantly higher abundances of *G. adiacens*, *S. intermedius*, and *M. tuberculosis*, in positive cases, with additional significant increases evident within *Streptococcus viridans* and *Mycobacterium bovis* abundances. No significant differences, between LC+ and LC- samples, within eukaryotic taxonomy were identified.

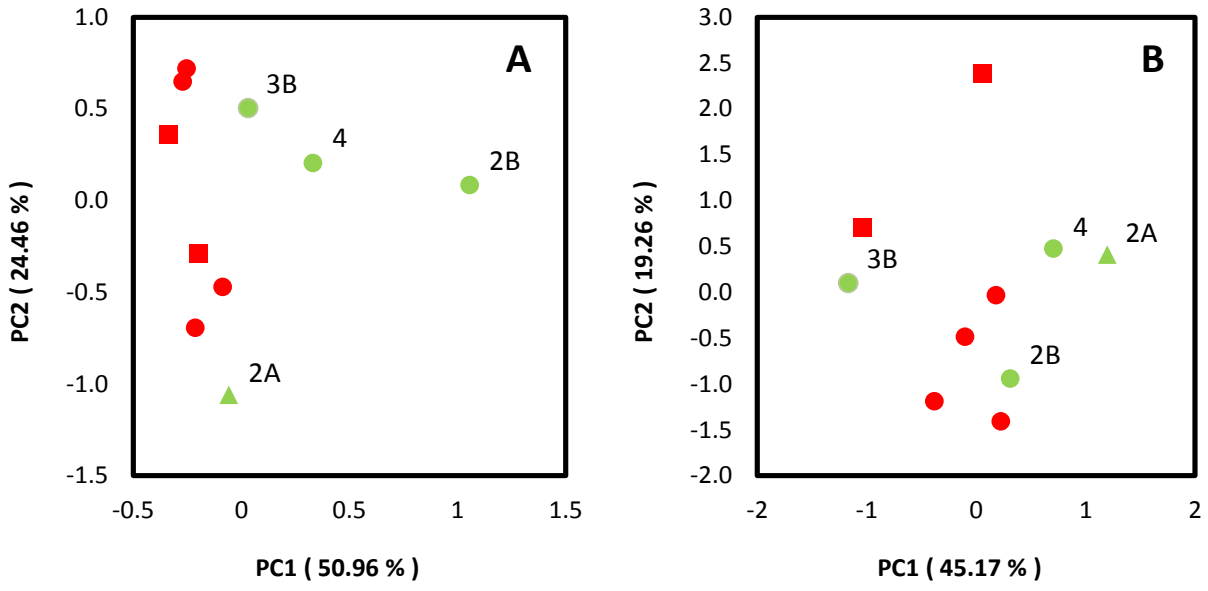


FIGURE 2.3 | Principal Component Analysis of Taxonomic and Functional Classifications
PCA plots created using the MG-RAST analysis platform, modelling (A) taxonomic, and (B) functional classifications, using the sequence cut-offs as detailed previously. PCA plots were drawn using normalised values and Maximum distance. Positive cancer samples are coloured green and negative coloured red. ■ symbol denotes an ex-smoker, ● a current smoker, and ▲ a never smoker. The code to the top-right of each lung cancer positive sample indicates the stage of the lung cancer at the time of sampling.

TABLE 2.3 | Average Percentage Abundance of Species Present in 'Core' Microbiome

Average percentage abundance of species present in the negative and positive lung cancer groups, with corresponding *P* values from one-way ANOVA, with significant values in bold text. % column shows average abundance, St. Dev. column shows standard deviation, and Count column shows the number of patients in each group in which the species was found, out of the total number of patients. The top division represents species present in all samples, the second division those found in all negative samples, and the bottom division those found in all positive samples.

Species	Lung Cancer Negative			Lung Cancer Positive			*P* Value
	%	St. Dev.	Count	%	St. Dev.	Count	
Streptococcus viridans	**0.02%**	**0.01%**	**6 / 6**	**0.10%**	**0.04%**	**4 / 4**	**0.042**
Streptococcus thermophilus	1.73%	1.33%	6 / 6	5.96%	2.51%	4 / 4	0.115
Ochrobactrum anthropi	0.48%	0.20%	6 / 6	0.95%	0.32%	4 / 4	0.188
Streptococcus pneumoniae	7.30%	1.84%	6 / 6	15.17%	6.72%	4 / 4	0.200
Enterococcus faecalis	0.10%	0.03%	6 / 6	0.25%	0.13%	4 / 4	0.203
Salmonella enterica	0.01%	0.00%	6 / 6	0.05%	0.04%	4 / 4	0.271
Neisseria meningitidis	1.04%	0.73%	6 / 6	0.79%	0.30%	4 / 4	0.756
Escherichia coli	0.01%	0.01%	6 / 6	0.06%	0.04%	3 / 4	0.207
Fusobacterium nucleatum	0.16%	0.14%	6 / 6	0.00%	0.00%	2 / 4	0.310
Haemophilus influenzae	22.29%	17.38%	6 / 6	4.30%	3.70%	3 / 4	0.346
Streptococcus parasanguinis	9.87%	4.24%	6 / 6	4.75%	4.75%	1 / 4	0.398
Streptococcus pyogenes	0.24%	0.11%	6 / 6	0.31%	0.12%	3 / 4	0.659
Veillonella parvula	1.60%	0.64%	6 / 6	2.01%	1.84%	2 / 4	0.805
Granulicatella adiacens	**0.00%**	**0.00%**	**1 / 6**	**0.07%**	**0.03%**	**4 / 4**	**0.015**
Streptococcus intermedius	**0.00%**	**0.01%**	**2 / 6**	**0.06%**	**0.02%**	**4 / 4**	**0.023**
Mycobacterium tuberculosis	**0.00%**	**0.00%**	**5 / 6**	**0.01%**	**0.00%**	**4 / 4**	**0.036**
Enterococcus sp. 130	0.01%	0.01%	5 / 6	0.06%	0.02%	4 / 4	0.050
Streptococcus sp. 6	0.01%	0.01%	5 / 6	0.06%	0.02%	4 / 4	0.050
Acinetobacter junii	0.01%	0.00%	5 / 6	0.03%	0.01%	4 / 4	0.112
Streptococcus salivarius	0.14%	0.10%	4 / 6	3.23%	2.46%	4 / 4	0.195
Shewanella spp.	0.00%	0.00%	3 / 6	0.01%	0.00%	4 / 4	0.206
Streptococcus vestibularis	0.01%	0.01%	5 / 6	1.31%	0.74%	4 / 4	0.208
Lactobacillus paracasei	0.00%	0.00%	1 / 6	2.61%	2.60%	4 / 4	0.234
uncultured bacterium	0.00%	0.00%	4 / 6	0.00%	0.00%	4 / 4	0.266
Lactococcus lactis	0.02%	0.01%	4 / 6	0.02%	0.01%	4 / 4	0.379
Neisseria gonorrhoeae	0.29%	0.20%	5 / 6	0.17%	0.08%	4 / 4	0.500
Rhodococcus erythropolis	0.09%	0.11%	3 / 6	0.13%	0.06%	4 / 4	0.628
Stenotrophomonas maltophilia	0.10%	0.05%	5 / 6	0.20%	0.08%	4 / 4	0.662
Staphylococcus aureus	0.42%	0.49%	5 / 6	0.20%	0.14%	4 / 4	0.666

2.3.4 | Comparison of Functional Capability of Microbiome

With regards to PCA separation of functional classifications, Figure 2.1b, no obvious separation was evident. This appeared to be reflected in both lung cancer status, and smoking status. Again, using MetaboAnalyst 2.0, significant fold changes in functional alignments were also identified. At the crudest level of functional classification, Level 1, no differences were evident. However, at Levels 2 and 3, Table 2.5, significant differences were observed. At Level 2, four functional classifications, involved in the urea cycle, putrescine/gamma-aminobutyric acid (GABA) utilisation, Gram-positive cell wall components, and invasion and intracellular resistance, were higher in LC+ samples. At Level 3, seven functional classifications were higher in LC+ samples, whilst three were lower, when compared to LC- samples. These differences appeared to be across a wide range of biological function, with higher levels of classifications observed for methicillin resistance in Staphylococci, iron sidrephore sensor and receptor system, amongst others. Lower levels of functional alignments for LC+ samples were observed in the COG0451 family, RNA methylation, and glutaredoxins.

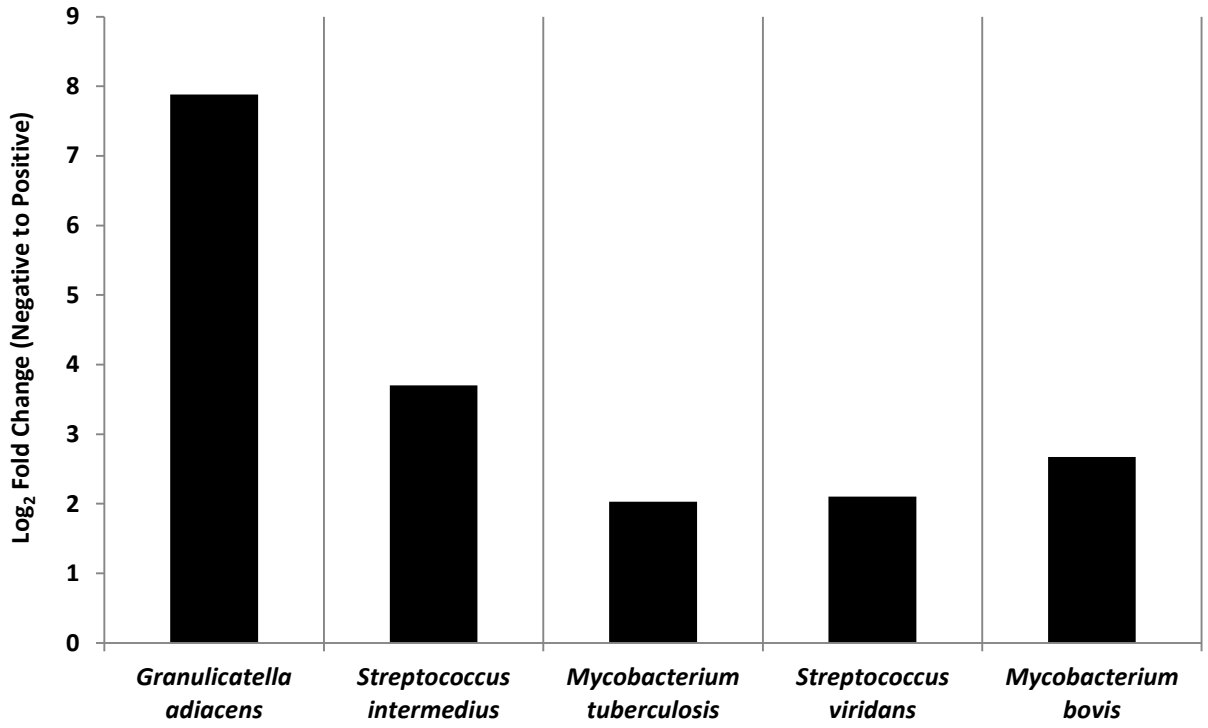


FIGURE 2.4 | Significant Fold Changes in Species Abundance from LC- to LC+
Using the online features of MetaboAnalyst 2.0, significant fold changes, as determined by t-Tests with *P* values <0.05, were identified. Five species, from three genera, were all higher in positive lung cancer samples, with *Granulicatella adiacens* and *Streptococcus intermedius* showing the highest change.

2.3.5 | Microbiomic Biomarkers for Lung Cancer

To evaluate the potential of using metagenomics to identify novel biomarkers for lung cancer and lung cancer progression, both a species level regression analysis and Level 3 functional regression analysis were completed. Those regressions with an R^2 value of 80% or more were plotted to identify those with differing relationships between negative and positive lung cancer groups. From this, the bacterial species *G. adiacens* was identified as having positive correlations with six other bacterial species, Figure 2.6, in positive lung cancer samples, which was not evident in in negative lung cancer samples. Additionally, when the stagings of positive lung cancer samples were plotted, a pattern with disease progression was observable, with a linear relationship seen whereby advances in lung cancer staging was matched by increases in *G. adiacens* level and the six other bacterial species detailed in Figure 2.6.

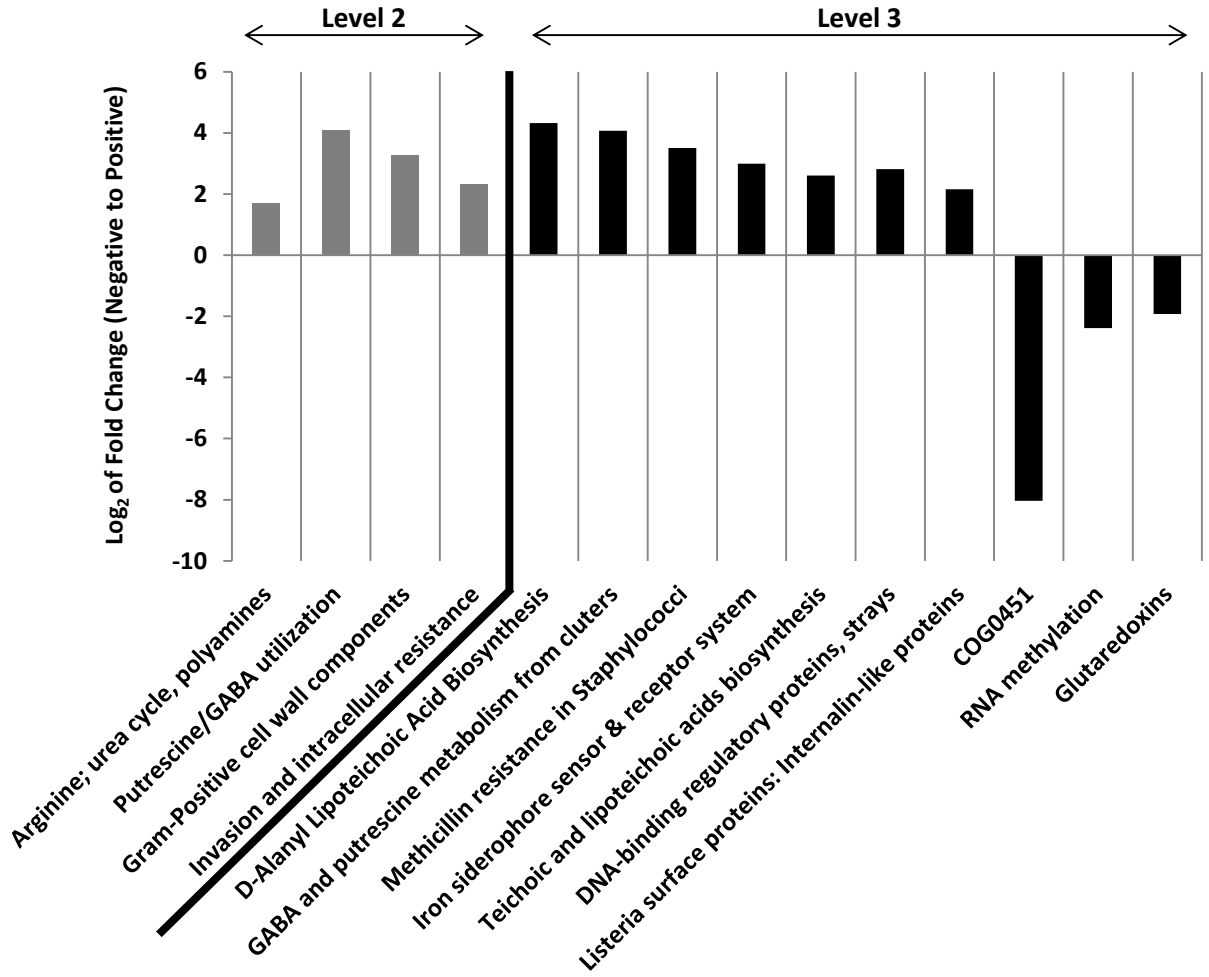


FIGURE 2.5 | Significant Fold Changes in Levels 2 and 3 Functions from LC- to LC+
Using MetaboAnalyst 2.0, significant fold changes of Level 2 (grey bars) and 3 (black bars) functional alignments, as determined through t-Tests with *P* values <0.05, were identified. A total of four Level 2 functional alignments were higher in positive lung cancer, alongside seven Level 3 functions. Three Level 3 functional alignments were lower in positive lung cancer samples.

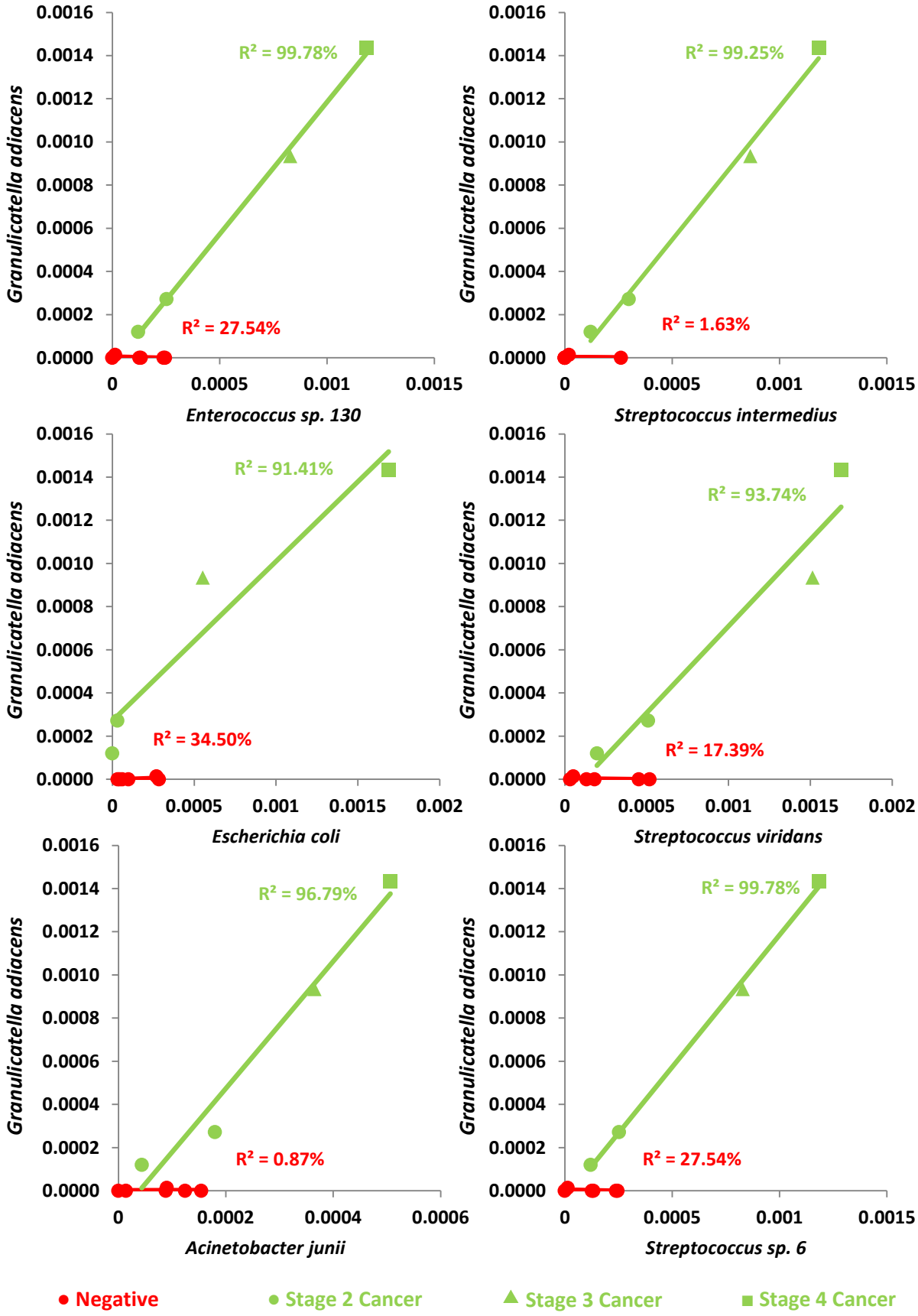


FIGURE 2.6 | Regression Analysis Suggests Importance of *G. adiacens* in LC+ Analysis
Species regression analyses were completed, and those with an R^2 value of greater than 80% were plotted to identify differing relationships between negative and positive lung cancers. This type of relationship was shown to exist between *G. adiacens* and six other species, with a strong positive relationship present in positive lung cancer samples, and no correlation evident within negative lung cancer samples. Normalised percentage abundances are shown on *x* and *y* axes.

2.3.6 | Patient Cohort for Lung Metabolome Study

Patients and participants sampled as part of this study are summarised in Table 2.4, with full clinical information detailed in Chapter 2 Appendix, Supplementary Table 2.3. A total of 34 clinical patients with suspected lung cancer were recruited, with 23 confirmed with lung cancer (LC+) (16 NSCLC (nine Stage 4, three Stage 3A, one Stage 3B, three Stage 2B, and one Stage 1B), six SCLC (three extensive and three limited), and one receiving a clinic-radiological diagnosis made by the lung cancer multidisciplinary team), and 11 had no diagnosis of lung cancer after extensive testing and follow up for at least one year (LC-). In addition, a total of 33 non-clinical controls were collected from staff and students at Swansea University, with no history of clinical lung disease. No discernible differences were observed between the information detailed in Table 2.4, except that within the LC- group there was a gender bias towards a significantly higher proportion of male patients than female patients.

TABLE 2.4 | Summarised Patient and Participant Information
Summarised patient information detailing clinical data. Full clinical data for clinically acquired samples, and information collected for healthy control participants, are fully detailed in Chapter 2 Appendix, Supplementary Table 2.3. nc = data not collected. Smoking pack years is individual history normalised to 20 cigarettes a day for one year.

	Non-Clinical Controls (CON)	LC Negative (LC-)	LC Positive (LC+)
Number	33	11	23
Age	55.3 (14.6)	66.5 (14.3)	66.6 (8.1)
Gender			
Male	20	10	11
Female	13	1	12
Smoking Status			
Current	15	3	10
Ex	0	8	10
Never	18	0	3
Smoking Pack Years	NC	49.0 (34.9)	39.3 (18.9)
Infection Present			
Yes	NC	3	1
No	NC	8	22
CO Level (ppm)	NC	3.7 (1.3)	4.2 (2.8)

2.3.7 | Comparison of LTQ-MS and GC-MS Metabolite Analysis

Preliminary analysis of metabolomic fingerprints acquired through LTQ-MS and GC-MS was conducted through principal component analysis. From this, the metabolomic fingerprints acquired in negative LTQ-MS mode, Figure 2.7a, showed the greatest degree of separation between the three samples groups. This separation was not evident in positive LTQ-MS mode, Figure 2.7b, and only partially present in GC-MS profiles, Figure 2.7c. Similar degrees of separation were seen in hierarchical cluster analysis,

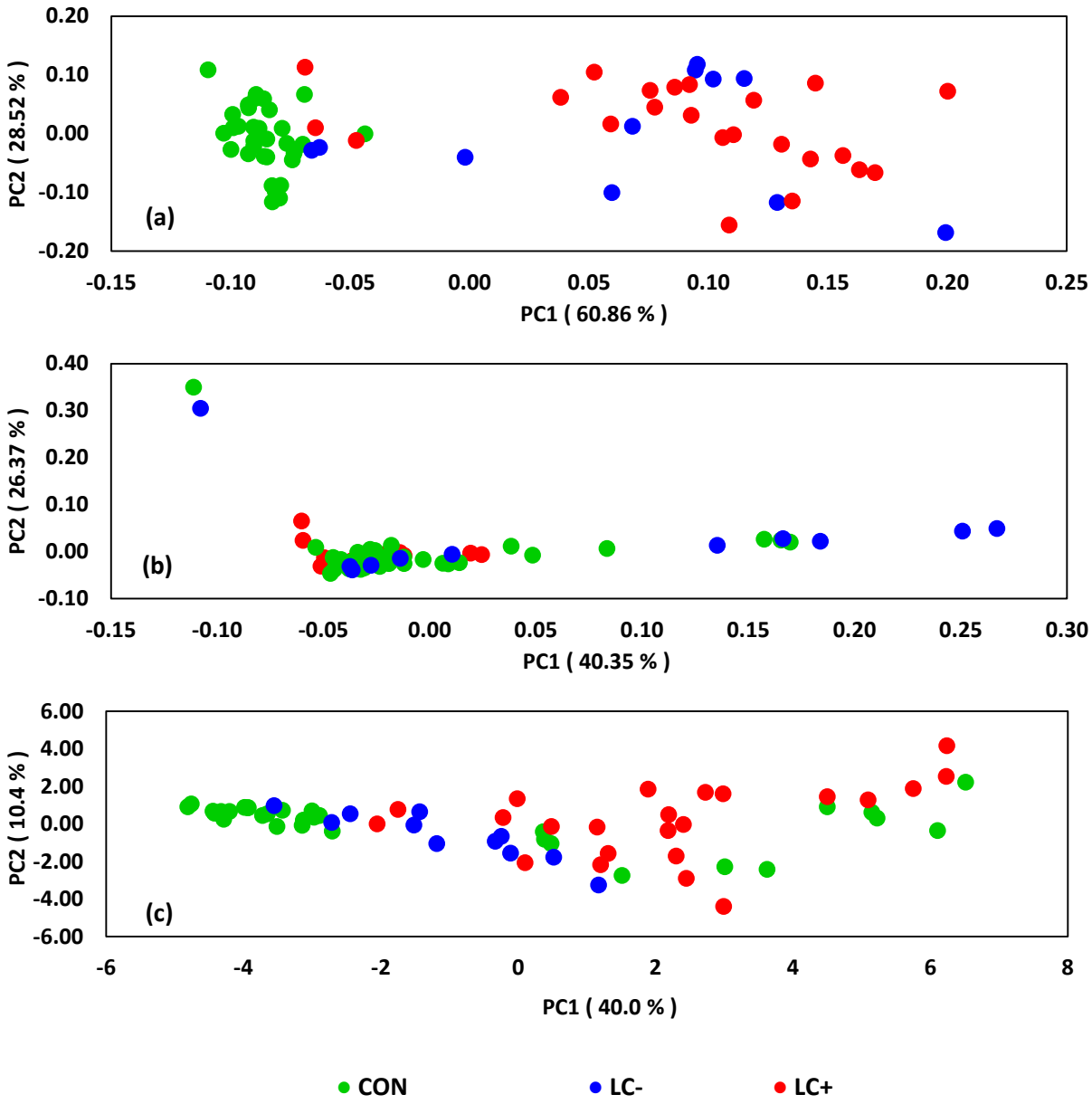


FIGURE 2.7 | Principal Component Analysis Plots for LTQ and GC-MS Fingerprinting
PCA, based on metabolites acquired in (a) LTQ-MS negative mode, (b) LTQ-MS positive mode, and (c) GC-MS, clearly differentiates between the clinically and non-clinically acquired samples, though separation of the two clinical groups, lung cancer and symptom controls, does not occur. For (c), individual plots are coordinate means of two duplicate samples.

based on the top 25 metabolites identified through one-way ANOVAs, LTQ-negative metabolites, Figure 2.8. Again, clear separation was evident between clinical and non-clinical samples, but not between LC+ and LC- groups.

2.3.8 | Identification of Clinically Relevant Biomarkers

Due to the greater degree of separation shown in negative mode LTQ-MS fingerprints, it was decided to use this for the identification of clinical relevant metabolomic biomarkers. Initial analysis was completed using MetaboAnalyst 2.0's random forest analysis feature. This showed a number of differential metabolites, in LTQ-negative mode, which were altered in LC+ samples, compared to LC- and CON samples. This suggested that a number of LTQ-MS metabolites may be useful as biomarkers of lung cancer status. Using the online facility ROCCET, univariate analysis was completed to identify potential biomarkers based on area under the receiver operating characteristic curve. The top three metabolites are listed in Table 2.5, with the area under the curve (AUC) figure for the top differential metabolites shown in Figure 2.10. This identified a number of metabolites that had a high AUC value (>0.99) for

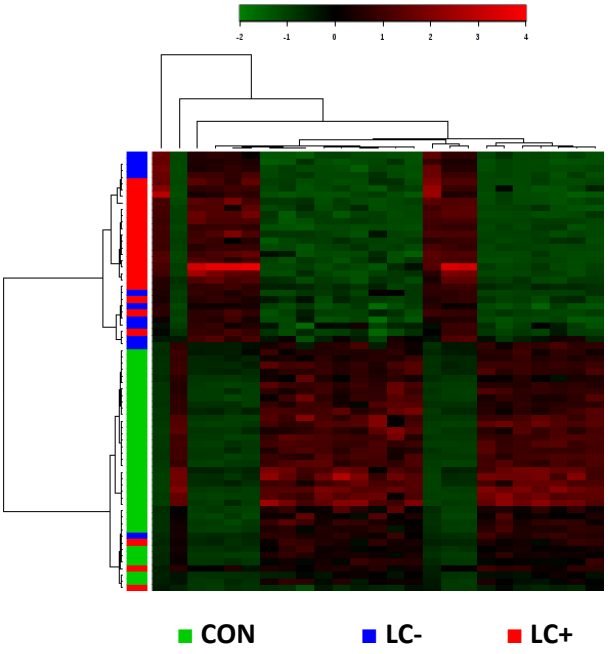


FIGURE 2.8 | Hierarchical Cluster Analysis with Heat Mapping for LTQ Metabolites
Hierarchical cluster analysis and corresponding heat maps were constructed, based on the top 25 metabolites identified through one-way ANOVAs, for metabolites identified in LTQ-negative mode. Similarly to PCA plots, separation between the clinically and non-clinically acquired samples was clear, but separation between LC positive and negative samples was not evident.

differentiating between clinically and non-clinically acquired samples. Additionally, a number of metabolites were identified with an AUC value of greater than 0.80, a threshold for clinically useful prediction. In LTQ-MS negative mode, metabolite 1496.72 showed the highest AUC value. At this value, 0.85, the metabolite had a sensitivity of 54.5% and a specificity of 69.6%.

TABLE 2.5 | Top Five Area Under Curve Values for LTQ-Negative Mode Metabolites

Using the online ROCCET platform, the top five metabolites, based on AUC values, for each differential group comparison were identified. For clinical and non-clinical comparisons, high AUC values were obtained, and for the LC negative and positive comparison, five metabolites with AUC values greater than 0.8 were identified.

Differential	Metabolite	AUC Value	t-Test	Fold Change
CON Vs LC-	67.18/1479.81	1.00	4.47×10^{-15}	-2.38
	67.18/1560.81	1.00	3.10×10^{-15}	-2.26
	69.09/1479.81	0.99	6.09×10^{-15}	-2.25
LC+ Vs CON	53.27/75.09	1.00	8.74×10^{-24}	4.71
	69.09/75.09	1.00	7.42×10^{-24}	4.69
	189.09	1.00	6.12×10^{-20}	2.57
LC+ Vs LC -	1496.72	0.85	2.93×10^{-3}	-0.03
	957.36	0.83	4.57×10^{-3}	0.31
	1382.45	0.83	5.93×10^{-4}	0.07

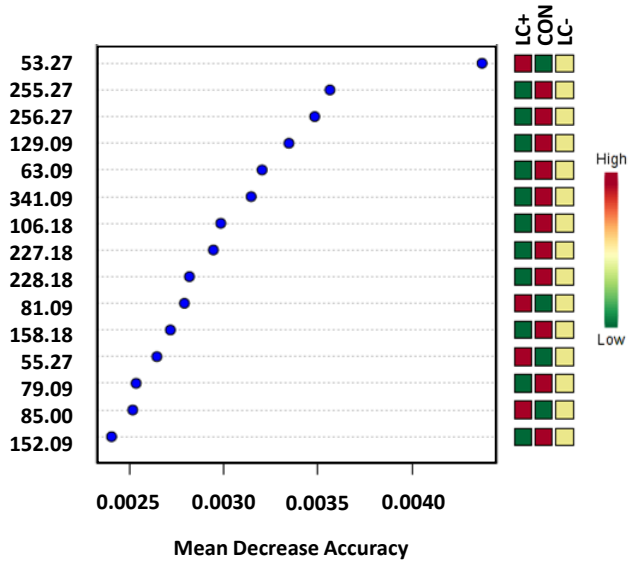


FIGURE 2.9 | Random Forest Plots for Identification of Key LTQ Metabolites

Random forests plots were constructed, using MetaboAnalyst 2.0 for LTQ-negative mode, which revealed a number of metabolites, in both modes, which may have potential in terms of diagnostic markers, particularly those that are either higher or lower in the LC+.

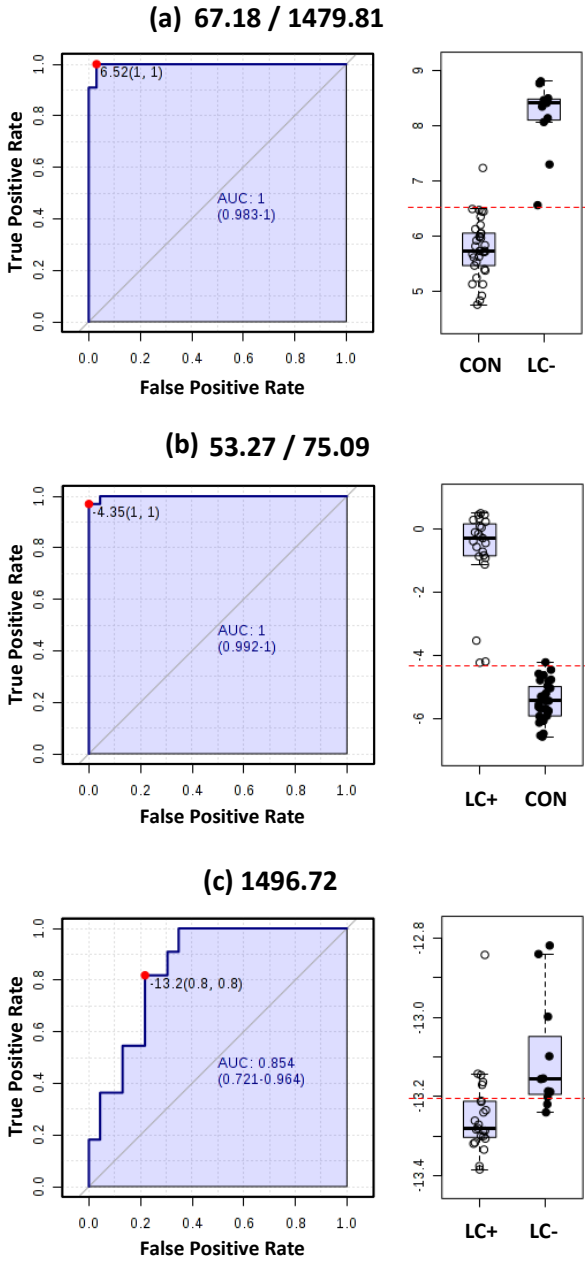


FIGURE 2.10 | Univariate Receiver Operating Characteristic Curve Analyses
Using the online facility, ROCCET, univariate receiver operating characteristic curves were created, and plotted to create area under the curve figures for LTQ-negative mode metabolites. The metabolite with the highest AUC value for each differential group is plotted.

2.4 | Discussion

Two separate approaches were explored with a view to identify novel biomarkers for lung cancer, namely exploring the lung microbiome and the lung metabolome. Overall, both approaches appear to have been successful in identifying biomarkers for lung cancer status, and staging in regards to the lung microbiome.

2.4.1 | The Lung Cancer Microbiome

The role of the microbiome in a range of respiratory conditions has been well documented; however, lung cancer has received only minimal attention. The lung cancer microbiome has been detailed, at the genus level, in female non-smokers from Xuanwei, China, through the use of amplicon sequencing. Interestingly, significant differences were only detected between sputum samples, and not buccal samples, suggesting a localised effect in the bronchial tree of the lung (Hosgood *et al.*, 2014). This study suggested a potential role of household coal burning exposure, and its effect on the lung microbiome in patients with lung cancer, rather than tobacco smoking, which is the most common cause of lung cancer in more economically developed countries (Jemal *et al.*, 2010a). This portion of work looked to address this, and to develop a more in-depth view of the microbiome with clinically relevant samples.

2.4.2 | Taxonomic Composition of the Lung Cancer Microbiome

In regards to PCA separation, Figure 2.3, taxonomy appeared to be influenced, to a greater degree than function, by lung cancer status. Interestingly, smoking status did not appear to be an influencing factor in separation. The effect of smoking on the lung microbiome is still to be firmly established (Erb-Downward *et al.*, 2011). To date, this is the first study to evaluate the lung cancer microbiome in a cohort of patients that contains smokers. It may be that any effect of smoking on the lung microbiome is outweighed by changes associated with lung disease, such as lung cancer.

The ‘core’, species-level microbiome, Table 2.3, revealed some noteworthy differences between the lung microbiomes of patients negative and positive for lung cancer. A significant difference was observed for *S. viridans*, which was found all patient samples, with a five-fold increase in LC+ patients. This bacterial species has been cultured from lung cancer patients in recent studies, but to date, it has not be identified as a differential bacteria for lung cancer state (Laroumagne *et al.*, 2013). Additionally, *Granulicatella adiacens*, *Streptococcus intermedius*, and *Mycobacterium tuberculosis* were identified as being significantly higher in LC+ patients, although they were only found in all LC+ samples and not all LC- samples. Interestingly, a history of tuberculosis has been associated with an increased risk of lung cancer (Shiels *et al.*, 2011), and a significantly shorter survival rate (Heuvers *et al.*, 2012). In this study, none of the patients had a history of tuberculosis, as this would have prevented their inclusion in the study. It may be that *M. tuberculosis* is a member of the commensal lung microbiome, and indeed, it is predicted that one third of the global population are carriers of latent tuberculosis (Korbel, Schneider and Schaible, 2008). Nevertheless, the significantly higher levels of *M. tuberculosis* observed in LC+ patients warrants further study. The significant differences observed in the ‘core’ microbiome, Table 2.3, are mirrored in the significant fold changes from the LC- to the LC+ group, Figure 2.4, except that *Mycobacterium bovis* is also significantly higher in LC+ samples.

In this study of the lung cancer microbiome, the focus has been on characterising the bacterial component. However, the eukaryotic component of the lung microbiome may also hold important and novel insights. The number of eukaryotic microbes in the lung microbiome is likely to be orders of magnitude lower than the bacterial component, but nevertheless, may harbour a number of opportunistic pathogens (Huffnagle and Noverr, 2013). In this study, no significant differences between the eukaryotic portions of the lung microbiome were detected between LC- and LC+ patients. However, a high number of alignments were removed because of their classification to the *Homo sapiens* genome, or other non-microbial eukaryotic genomes. This is likely to represent a lack of high-quality databases for eukaryotic microbes, which is a severe limitation.

2.4.3 | Functional Capacity of the Microbiome

Metagenomic sequencing allows the field of microbiomics to move beyond characterising the microbiome simply in terms of its taxonomic composition, and more towards understanding how its functional capability shifts in response to disease state. Here a total of four Level 2 classifications, were identified that showed significantly higher levels in patients positive for lung cancer, including those involved in arginine use, urea cycle, putrescine and gamma-aminobutyric acid utilisation, and invasion and intracellular resistance. Interestingly, elevated levels of polyamines, such as putrescine and GABA, have been associated with a range of cancers including lung malignancies (Nowotarski, Woster and Casero, 2013). Polyamines offer a rich nitrogen source of bacteria, and elevated levels associated with lung cancer could explain why there are significantly more associated alignments in positive lung cancer cases.

At Level 3 of functional classification, seven functions were significantly higher, and three significantly lower, in the positive lung cancers. Some of these increases reflected Level 2 changes, but others, such as higher levels of iron siderophore sensors and receptor system alignments, further suggest that changes in the cancerous lung are reflected by changes in the lung microbiome. Elevated iron levels are associated with lung cancer (Xiong, Wang and Yu, 2014), and as iron is essential for many cellular functions in bacteria, it is not unexpected that elevated levels of iron associated with lung cancer would result in a selective pressure to reflect this in the microbiome. Functional changes are not unexpected in the lung microbiome as a response to malignancy formation, but this study is the first to confirm that such synergy exists.

Of the three functional classifications shown to be significantly lower in LC+ cases, two have known functions. The biggest reduction in functional classifications was seen in the COG0451 grouping, but to date, these have not been assigned a biological function. However, significantly lower levels were also seen in functional classifications with known biological function namely RNA methylation and glutaredoxins, although the significance of these changes, in regards to lung cancer, is unclear.

2.4.4 | The Microbiome as a Source of Novel Biomarkers

One of the key purposes of this portion of work was to identify potential novel biomarkers for lung cancer state and stage. To this end, *G. adiacens* was shown to have a significant positive relationship with six other bacterial species, which is only seen in patients positive for lung cancer. The *Granulicatella* genus has been identified as being significantly higher in the sputum of non-smoking lung cancer cases (Hosgood *et al.*, 2014), suggesting that it may be a true reflection of lung cancer state, rather than a by-product of tobacco smoking. The *Granulicatella* genus, and *G. adiacens* specifically, is a difficult organism to culture, which may be the limiting factor that explains the minimal study that has been conducted into it (Woo, 2003). It has however, been associated with endocarditis (Perkins *et al.*, 2003) and septicaemia (Bizzarro *et al.*, 2011). *G. adiacens* may be an opportunistic human pathogen, or a commensal bacterium that is able to take advantage of a niche, such as altered sputum composition (Hubers *et al.*, 2013), that is present in the lungs of patients with lung cancer. Additionally, the six other bacterial species associated with *G. adiacens* in positive lung cancer patients; do not appear to have been linked to lung cancer in the literature. Similarly, it may be that they respond to a selective pressure in the cancerous lung, or potentially, that *G. adiacens* has a synergistic relationship which enables their higher abundance. Regardless of the biological basis for these significantly higher abundances seen, they nevertheless have the potential to act as biomarkers for lung cancer, in regards to both lung cancer status, but also in staging, due to the pattern of abundances seen in Figure 2.6.

The use of spontaneous sputum is a well-established diagnostic medium for lung cancer because of its non-invasive collection, and that it is symptomatic of lung cancer as a disease. Therefore, it offers a viable alternative to radiography based diagnoses for high-throughput, non-invasive, and low-risk screens (D'Urso *et al.*, 2013). However, it should be appreciated that sputum production is localised to the upper respiratory tract, and particularly the bronchial tree. As microbiome studies in other respiratory diseases have shown, including in COPD (Erb-Downward *et al.*, 2011) and cystic fibrosis (Willner *et al.*, 2011), spatial differences can exist within the lungs and therefore, sputum should only be taken as representative of the microbiome in the upper respiratory tract.

This novel study has expanded upon the knowledge of the microbiome in patients with lung cancer, using clinically relevant control samples, particularly in regards to the functional capacity of the microbiome, and its taxonomic composition at the species level. Additionally, the strength of using metagenomics to identify potential biomarkers for disease state and progression has been demonstrated; namely *G. adiacens* abundance correlations with a range of bacterial species, which could have clinical use. However, due to the relatively small sample number in this pilot study, more work is needed to confirm these relationships, and whether they are observable in earlier stage lung cancers, and other types, such as SCLC.

2.4.5 | The Lung Cancer Metabolome

Since the 'Warburg Effect', as it is now know, was first described in 1956 (Warburg, 1956), the alterations that cells undergo during carcinogenesis has been a focus of both basic and applied clinical research. The concept that these cellular changes could be detected via the subsequent changes in low molecular weight compounds has received some attention in the field of lung cancer diagnostics and biomarker discovery. For example, GC-MS has been used to identify potential biomarkers for a range of cancers in the form of urinary volatile organic metabolites (Silva, Passos and Câmara, 2011). Saliva has also been used as a biofluid for metabolomics, with capillary electrophoresis mass spectrometry being used to identify specific oral, breast and pancreatic cancer profiles (Sugimoto *et al.*, 2010). Additionally, lung cancer has received attention in terms of metabolomics, but these investigations have focussed on the cancerous tumours themselves or serum from affected patients, and using a limited range of mass spectrometry techniques (Hori *et al.*, 2011).

In this exploratory study, sputum has been suggested as a source of biomarkers for the identification of positive or negative lung cancer status of a patient group presenting with lung cancer like symptoms. Through the use of ROCCET analysis, a number of negative mode LTQ-MS metabolites that had AUC values greater than 0.8, a cut-off for discrimination that may be useful in a clinical setting, were identified. The LC+ group consisted of a range of lung cancer stages and histology, including NSCLC and

SCLC. This suggests that biomarkers established through metabolomic techniques could have utility as a preliminary screen for lung cancer status, identifying patients that require a clinical follow-up for lung cancer confirmation, histology and staging.

2.4.6 | Metabolomic Biomarkers for Lung Cancer

As far as it is possible to ascertain, this is the first study to report on the metabolomic fingerprinting of sputum acquired from lung cancer patients. The use of sputum, the production of which is a symptom of lung cancer, as a biofluid for screening carries the benefit of being non-invasive, high-throughput, and low-cost, compared to current conventional methods such as CT scans (Rivera, Mehta and Wahidi, 2013).

Although the metabolites, with AUC values that suggest their potential clinical usefulness, identified in this portion of work, have a false positive rate that would prove problematic in a whole-population screen, they could still have utility in the screening of a 'high-risk' group, such as smokers. Furthermore, it may be that metabolomic biomarkers could be combined with other forms of biomarkers, such as circulating miRNAs, to enable an integrated approach to lung cancer screening, as has been suggested for other cancers (Laxman *et al.*, 2008).

In this portion of work, three different sample groups were used. In terms of a lung cancer screening regiment, it is likely that only ever an 'at risk' group of patients would be screened, such as smokers. The Control group in this work all have a history of smoking, albeit not all current, which would place them at an elevated risk of lung cancer. Here, a number of metabolites, or metabolite ratios, Table 2.5, have been shown to have an AUC value of 1.0 between the Control and LC+ group. This would translate into sensitivity and specificity percentages of nearly 100%. The main technology currently being proposed for lung cancer screening is CT scanning, which does not have sensitivity and specificity as high as shown here, and also carries the issue of potential risk to patients through repeated radiation exposure (Bach *et al.*, 2012).

One of the strengths of this portion of work is the use of patient samples from a group presenting with suspected lung cancer, but not all of whom went on to be positively diagnosed with the disease. This means that metabolomic biomarkers for lung cancer that were able to differentiate between these two patient groups would have clinical use as a preliminary, non-invasive screen. This could, potentially, then identify patients who would need follow-up assessment with CT scanning, or another diagnostic such as bronchoscopy. Low-dose computed tomography scans are widely considered to be the most effective method of screening for lung cancer. However, the cost of LDCT screening currently prevents its widespread adoption, with each false-positive estimating to cost $1,000 (Wood *et al.*, 2012). The use of metabolomic biomarkers, as described in this portion of work, could be employed to reduce the number of patients referred for LDCT, the consequent number of false-positive diagnoses, and thus the cost of screening using LDCT.

2.4.7 | Moving from Fingerprint, to Metabolome, to Metabolite

In this portion of work, a metabolomic fingerprinting method was employed to identify potential metabolomic biomarkers. However, as previously discussed, metabolomic fingerprinting identifies biomarkers without identifying individual metabolites. Although this provides for a fast classification of samples based on mass spectrometry, or additionally NMR, it does not provide insight into the biological processes involved in altering the metabolome to such an extent that allows for separation between disease groups, as seen here (Madsen, Lundstedt and Trygg, 2010). To progress this work further, identification of individual metabolites responsible for the high AUC values observed would be required. This could be accomplished through the use of a variety of mass spectrometry method. After the identification of individual metabolites, the biological pathways that each are involved in could be further interrogated to understand aberrations causing a detectable change in the metabolome of specific disease states, such as lung cancer presence. The identification of individual metabolites would also improve a screening methodology based on a mass spectrometry-based approach. For example, if a small number of metabolites have been identified, such as that been demonstrated in breast and ovarian cancer (Slupsky *et al.*, 2010).

The utilisation of metabolomic fingerprinting to identify potential biomarkers for the identification of lung cancer status is novel and promising. However, there are a number of limitations to this portion of work. Firstly, the sample size used is relatively small, with minimal early stage lung cancer samples. This means that accurate predictions for the applicability of the findings to this group are not possible. Furthermore, due to insufficient sputum volume, accurate identification of the biomarkers with potential clinical application was not possible. Nevertheless, the power of using metabolomics to identify biomarkers, with potential clinical application, to separate those presenting with suspected lung cancer, into those positive and negative for the disease, has been shown.

2.8 | Conclusions and Future Work

In this portion of work, the lung microbiome and metabolome were investigated to evaluate them as a source of novel biomarkers for the identification of lung cancer. To this end, the lung cancer microbiome and metabolome have been shown to possess unique characteristics, when compared to those without lung cancer, which are sufficient to differentiate disease groups.

For the lung microbiome, a number of bacterial species were identified as being potential biomarkers for both lung cancer status and stage. Additionally, for the metabolome, a number of fingerprinting metabolites were shown to be able to differentiate between lung cancer positive and lung cancer negative patients, to a level that may have clinical usefulness. Both of these approaches employed sputum as a sampling biofluid. This is arguably one of the main strengths of both of these portions of work because sampling sputum is non-invasive. Furthermore, because it is localised to the lungs, and specifically the bronchial tract, the aberrations caused by the presence of lung cancer are likely to be concentrated, unlike other biofluids such as blood or urine, where the changes may be diluted to such an extent that they are not detectable.

An additional strength of this portion of work is the patient cohort that was investigated. Albeit a small, pilot-sized study, the use of patients all presenting with suspected lung cancer, only a portion of whom were subsequently diagnosed with lung cancer, means that the changes observed in the lung microbiome and metabolome are not simply the result of general lung disorder.

Both the lung microbiome and metabolome portions of work presented here have revealed novel insights into lung cancer. Nevertheless, the are both relatively small-scale pilot studies, and future work, with a considerably increased cohort of patients, with a variety of lung cancer stages and histologies, will be required to verify the findings presented here. Additionally, this presents the opportunity to combine microbiome and metabolome investigations towards a systems biology approach for the identification of lung cancer biomarkers.

CHAPTER 3 | Understanding the COPD Microbiome Through Metagenomic Sequencing

CHAPTER SUMMARY | The lung microbiome in COPD has been well-characterised but little is known about the functional capacity of the microbiome, which can be revealed through metagenomics. Genomic DNA was isolated from spontaneous sputa (sampling the upper respiratory tract) from ten control participants and eight patients with moderate (three Global Initiative for Chronic Obstructive Lung Disease (GOLD) II) to severe COPD (five GOLD III). Nextera® libraries were constructed and sequenced on the Illumina HiSeq 2500 platform. Resulting sequences were then analysed using the MG-RAST pipeline; an automated analysis platform for metagenomes. Principal component analyses, irrespective of smoking status, showed partial separation at the level of bacterial taxon but only to a lesser extent based on function. A core microbiome of eight microbial genera (*Haemophilus, Lactobacillus, Neisseria, Ochrobactrum, Pseudomonas, Staphylococcus, Streptococcus,* and *Veillonella*) and four species (*Haemophilus influenzae*, *Ochrobactrum anthropic*, *Streptococcus pneumoniae*, and *Streptococcus thermophilus*) were found in all 18 samples. Significant differences were observed in the abundance of bacterial species between Control and COPD patients, particularly in regards to species of the *Streptococcus* genus. A total of 14 bacterial species were shown to have higher abundance levels in COPD patients, and a total of 17 bacterial species were shown to have lower abundance levels in COPD patients. In addition, functional differences in the metagenomes from COPD patients were consistent with a greater bacterial growth capacity, with higher abundances of genes involved in amino acid, carbohydrate, protein, DNA, and RNA metabolism. Regression analyses correlated COPD severity (as indicated by worsening airflow obstruction / lower FEV_1% of predicted) with differences in the *Streptococcus* genus, specifically *S. pneumonia*, and also functional classifications aligned to sialic acid metabolism. This pilot study has linked COPD severity to the abundance of *S. pneumonia* and altered bacterial sialic acid metabolism, which could influence the inflammatory response in the lung. Thus, the functional capacity of the COPD microbiome may be a factor in bacterial-associated exacerbations and raises a number of avenues for further study.

3.1 | Introduction

Chronic obstructive pulmonary disease is a leading cause of death and morbidity, leading to an estimated 2.75 million deaths worldwide in 2006 (Lopez *et al.*, 2006). COPD is usually caused by smoking in the developed world and is an umbrella term for a multisystemic inflammatory state, including several diseases such as chronic bronchitis and emphysema.

3.1.1 | COPD Aetiology and Pathogenesis

COPD does not have a single definitive cause, but rather a collection of risk factors. Tobacco smoking and certain occupational exposures are the two certain environmental risk factors and α_1-antitrypsin deficiency is a certain host risk factor. COPD remains as a poorly characterized disease, though the central principle is that lung function is limited through a combination of small airways disease and parenchymal destruction (Mannino and Buist, 2007).

Tobacco smoking is likely to be responsible for the vast majority of COPD cases worldwide, but there are noteworthy differences. For example, in less economically developed countries, a more important risk factor for COPD may be household biomass fuel smoke exposure. Indeed, with tobacco consumption decreasing in more economically developed countries, Figure 3.1, exposure to household smoke may be one of the more significant risk factors globally (Salvi and Barnes, 2010).

Conservative estimates of disease burden suggest that at least 10% of the adult population over the age of 40 years is affected by COPD. Additionally, the link between tobacco and COPD is reinforced by the considerably higher prevalence in current and ex-smokers compared to non-smokers. COPD prevalence is also increased in men, compared to women, which may be somewhat explained by the pattern of tobacco consumption described in Figure 3.1 (Halbert *et al.*, 2006).

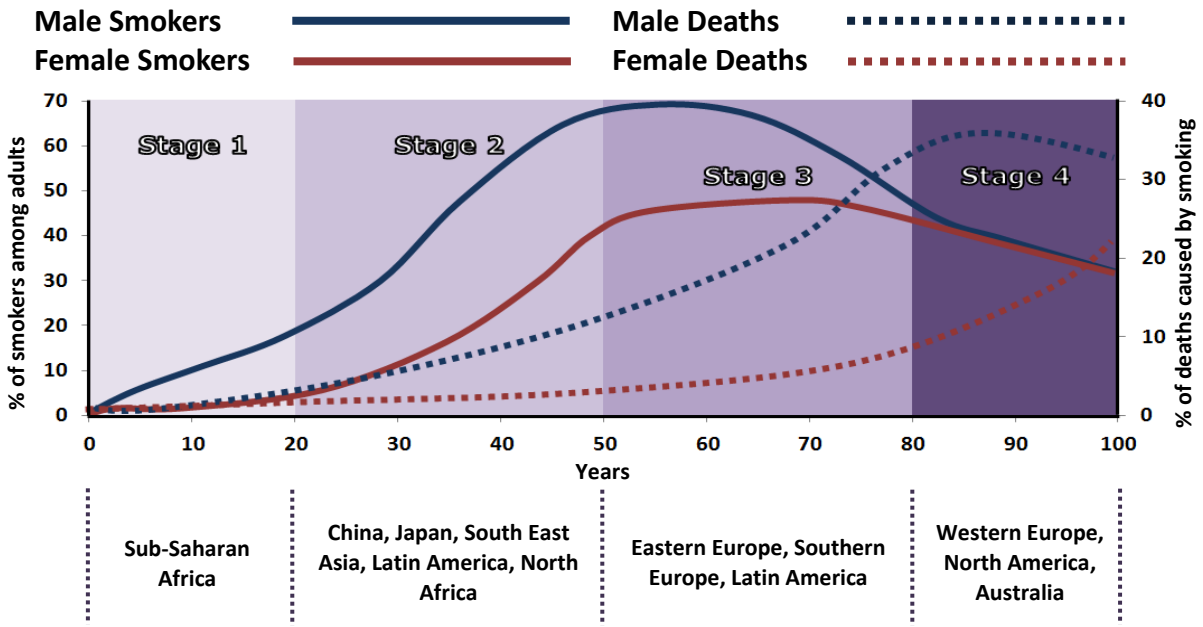


FIGURE 3.1 | Model of the Worldwide Tobacco Epidemic
This model describes the initial rise and subsequent decline in smoking prevalence which is followed, albeit with some delay, by a rise and then decline in the number of deaths caused by smoking. With this pattern of tobacco consumption worldwide, it is possible that tobacco smoking may cease to be the primary cause of COPD. Figure adapted from Lopez, Collishaw and Piha, (1994).

The morbidity of COPD is one of the primary social and economic burdens of the disease. Measures of morbidity traditionally cover visits to primary and secondary healthcare, number of prescriptions, and overnight hospital admissions. Although exact figures for COPD morbidity are difficult to calculate, because of the numerous comorbidities associated with the disease, it is established that COPD-related morbidity increases with age (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014). In addition to the significant mortaility associated with COPD, with recent modelling predicting that COPD will be the fourth leading cause of mortality worldwide by 2030 (Mathers and Loncar, 2006), the economic burden of COPD makes it a significant global issue.

The pathogenesis of COPD is defined primarily by the inflammatory response of the respiratory tract, which is modified in the disease. This in turn is amplified by subsequent exposure to irritants, including tobacco and household fuel smoke, which is sustained after exposure is ceased; making COPD an irreversible disease. A number of factors can be involved in modification of the inflammatory response, including constituents of the lung microbiome, oxidative stress, increased protease levels, and inflammatory cells and mediators (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014).

One of the key characteristic of COPD is periods of acute symptom exacerbations, leading to a substantial worsening of morbidity and mortality. It is defined as a natural event in COPD which is characterized by a change in a patient's baseline dyspnea, cough, and/or sputum production, which is beyond normal day-to-day variations and may warrant a change in regular medication (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014). Increased levels of inflammation are very common in acute exacerbations, and are usually triggered by colonisation of the lower respiratory tract by viruses, bacteria, or a combination of both (Wedzicha and Seemungal, 2007). The level of bacterial load in COPD acute exacerbations has been shown to be increased compared to baseline COPD, suggesting that the microbiome may play a significant role in the initiation of exacerbations, immune response to exacerbations, or effectiveness of clinical interventions (Garcha *et al.*, 2012). Therefore, establishing the COPD microbiome at a baseline of the disease may have important implications in improving understanding of acute COPD exacerbations, and suggesting potential predictors for them.

3.1.2 | Diagnosis and Treatment of COPD

A clinical diagnosis of COPD is usually considered when a patient presents with dyspnea, a chronic cough

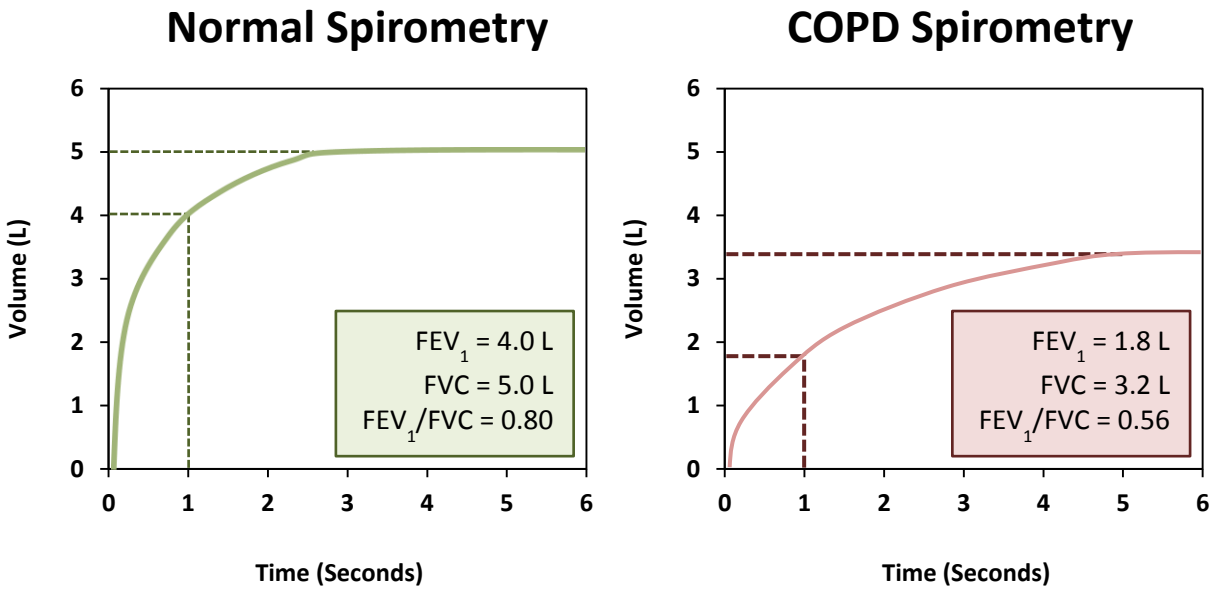


FIGURE 3.2 | FEV_1/FVC Spirometry for Normal and COPD-Affected Lungs
The use of FEV_1/FVC spirometry is required for a definitive clinical diagnosis of COPD. A FEV_1/FVC of less than 0.70 confirms the presence of COPD. Figure adapted from Global Initiative for Chronic Obstructive Lung Disease (GOLD), (2014).

or sputum production, and a history of exposure to COPD risk factors, such as tobacco smoking. A definitive clinical diagnosis requires spirometry, the measurement of breath, assessment. A post-bronchodilator forced expiratory volume in one second (FEV_1) over forced vital capacity (FVC) ratio, also defined as FEV_1% of predicted, the proportion of a patient's lung capacity that they are able to expel within one second (FEV_1) as a proportion of the total air that can be expelled from the lungs after full inspiration (FVC), Figure 3.2, of less than 0.70 confirms the presence of COPD (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014).

In addition to clinical diagnosis, spirometry is also used to classify COPD, based on the FEV_1/FVC ratio, from mild to very severe COPD, Table 3.1. These four COPD classifications carry distinct clinical characteristics, in terms of the predicted number of exacerbations and hospitalisations per year, and the three-year predicted mortality rate (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014). Although spirometry is the main method of classifying COPD, a weak correlation between the FEV_1/FVC ratio a patient's health-related quality of life has been shown. This means that within any of the four GOLD classifications of COPD, there is a significant degree of individual variation, with two patients sharing the same FEV_1/FVC ratio potentially having vastly different health-related qualities of life. This suggests that the comorbidities commonly associated with COPD also need to be formally

TABLE 3.1 | Classification of COPD with Clinical Characteristics of Groups
COPD is classified into four Global Initiative for Chronic Obstructive Lung Disease (GOLD) categories from mild to very severe COPD. This is determined by FEV1/FVC ratio. Clinical characteristics for GOLD I (Mild) COPD classifications have yet to be determined. Table constructed using data taken from Global Initiative for Chronic Obstructive Lung Disease (GOLD), (2014).

GOLD Rating	Description	FEV_1/FVC Ratio	Annual Exacerbations	Annual Hospitalisations	Three Year Mortality Rate
I	Mild	≥ 0.80	?	?	?
II	Moderate	≤ 0.50 to < 0.80	0.70 – 0.90	0.11 – 0.20	11%
III	Severe	≤ 0.30 to < 0.50	1.10 – 1.30	0.25 – 0.30	15%
IV	Very Severe	< 0.30	1.20 – 2.00	0.40 – 0.54	24%

assessed, as these may also have a significant influence on a patient's health-related quality of life (Jones, 2009).

One of the key treatment options for COPD patients is smoking cessation. This has been shown to be the intervention with the greatest degree of benefit in regards to moderating the progression of COPD. This benefit is likely brought about through the reduction in irritant exposure, namely cigarette smoke, which reduces inflammatory-related damage to the lungs. To date, there has been no medication for COPD that has been proven to either prevent, or reverse, the long-term decline in lung function that is characteristic of COPD (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014).

Because no treatment for COPD, to date, has been proven to be effective at improving lung function, treatments focus on the management of the symptomatic load of a patient. For stable COPD, there are two approaches treatment can take: pharmacological and non-pharmacological therapies. For pharmacological therapies bronchodilators such as β_2-agonists, anticholinergics, and corticosteroids are the main options. The overriding issue in COPD treatment is tailoring the regiment to individual patients. There is a significant degree of individual differences within COPD patients, particularly in regards to comorbidities, so these must be taken into consideration when treatment options are decided. Complications of pharmacological therapies can be significant, with increased risk of pneumonia for corticosteroid use for example. Non-pharmacological treatments for COPD can include pulmonary rehabilitation, oxygen therapy, or lung transplantation (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014).

The treatment of acute exacerbations is one of the most challenging areas in managing COPD. What constitutes an exacerbation is still poorly defined, and diagnosis and treatment relies on the patient presenting themselves with an acute change in their day-to-day symptoms. The majority of COPD acute exacerbations are believed to have a viral or bacterial cause, or a combination of both, with a limited number of microbial species believed to be the causative factor in the majority of acute exacerbations.

Antibiotic treatment has been shown to improve management of acute exacerbations, with the benefit enhanced in patients with severe COPD compared to those with moderate (Wedzicha and Seemungal, 2007). The clinical benefit of using antibiotics in the treatment of acute exacerbations of COPD suggests a significant role for the lung microbiome. Understanding the lung microbiome in stable COPD may reveal novel insights into the processes preceding acute exacerbations.

3.1.3 | The Lung Microbiome of COPD

It is now clearly established that the lungs are not a sterile environment (Charlson *et al.*, 2011). However, although much is known about the microbiomes of lungs in CF and asthmatic patients, relatively little is known about the lung microbiome of COPD patients. For example, it has not been established whether any changes in the lung microbiome of patients with COPD is a causative factor in the disease, or simply a reflection of the altered lung environment. Additionally, the role that common pharmaceutical treatments for the disease, such as inhaled corticosteroids, have on the lung microbiome have not been defined. Furthermore, because smoking tobacco is a major risk factor in developing COPD, the lung microbiome of "healthy" smokers also needs to be established.

To this end, Erb-Downward *et al*., used 16S rRNA pyrosequencing analysis to examine four participants with COPD, seven "healthy" smokers and three participants who have never smoked. Although the study did not find a significant difference in lung microbiomes between the three groups, it did conclude with three main results. Firstly, that the lung microbiome of "healthy" smokers is distinct to the microbiome reported for either the oral cavity or nasopharynx. Secondly, the diversity of the lung microbiome is less reduced in participants with decreased lung function, such as COPD, and is usually associated with *Pseudomonas* species. Lastly, the study found that some smokers had a less diverse lung microbiome relative to smokers with normal lung function which suggests that changes in the lung microbiota may be detectable before there is clinical evidence of reduced lung function via spirometric techniques (Erb-Downward *et al*., 2011).

As with CF, studying the lung microbiome of COPD patients may also aid in treatment in addition to its potential application in diagnosis. Acute exacerbations of COPD are a significant source of morbidity and in 50% of cases, bacterial infections are implicated. Presently, a small number of pathogens have been reliably identified in COPD airways, via culture-dependent methods, but the entire community structure of the COPD lung microbiome in acute exacerbations is still to be elucidated. Towards this goal, Huang *et al*., used 16S rRNA PhyloChip technology on a cohort of eight COPD patients in an attempt to identify significant pathogenic bacteria. They found that 75 taxa of bacteria were detectable in all COPD patients, and many of these taxa are known pathogens. Furthermore, they identified pathogenic bacteria, even in those patients who were currently undergoing antibiotic treatment for an infection. This suggests that improved treatment of exacerbations may be achieved through high-throughput identification of pathogenic bacteria to further identify appropriate therapeutic targets (Huang *et al*., 2010).

Patients experience acute exacerbations, which may be triggered by infection or environmental pollutants. Approximately 75% of acute exacerbations are attributed to viral or bacterial infection, or a combination of both (Han *et al.*, 2012). Characterisation of bacteria cultured from the airways of COPD patients has linked exacerbations with pathogens such as *Streptococcus pneumoniae*, *Haemophilus influenzae,* and *Moraxella catarrhalis* (Hirschmann, 2000).

Since the advent of culture-independent techniques, especially amplification and sequencing of the 16S rRNA gene, the lung microbiome of COPD has been one of the most widely studied amongst respiratory diseases. Analysis of the lung microbiome may lead to novel insights into the progression of COPD and potential therapeutic interventions. Lung microbiomic analyses based on 16S rRNA amplicon sequencing, to characterise bacterial taxonomic composition, compared bronchial alveolar lavages from patients with COPD to healthy individuals. Initial studies suggested that the lung microbiome of patients with moderate and severe COPD patients is less diverse than 'healthy' controls (Erb-Downward *et al.*, 2011), although other work suggested this was an underestimation of bacterial diversity (Huang *et al.*,

2010). More recent work, with a larger cohort of moderate and severe COPD patients, suggested increased microbial diversity in more severe COPD (Pragman *et al.*, 2012). Others have described a core COPD lung microbiome that includes *Pseudomonas, Streptococcus, Prevotella, Fusobacterium, Haemophilus, Veillonella*, and the *Porphyromonas* genera.

3.1.4 | Aims and Objectives of Chapter

COPD is one of the respiratory diseases that have received the greatest degree of attention in regards to analysis of the lung microbiome. However, metagenomic analysis of the lung microbiome of stable COPD patients has been distinctly limited. The aims and objectives of this portion of work were therefore, to:

1) Analyse the taxonomic make-up of the COPD microbiome, in stable patients, to the species level, compared to individuals without COPD.

2) Evaluate the functional capacity of the lung microbiome in stable COPD patients to identify potential implications that this may have on disease progression, or management.

3) Investigate characteristics of the lung microbiome in stable COPD patients to identify any that can be correlated with the severity of the disease.

3.2 | Materials and Methods

The MedLung study received loco-regional ethical approval (05/WMW01/75). Informed consent was obtained from all participants at least 24 hours prior to sampling, and at a previous clinical appointment for clinical patients. All data was link anonymised prior to analysis.

3.2.1 | Patient Recruitment and Sampling

Spontaneous sputum was collected from patients with a clinical and spirometric diagnosis of COPD from two UK hospitals. COPD diagnosis required a history of at least ten smoking pack years, an age of older than 40 years, and post bronchodilator FEV_1/FVC of less than 0.70. Additionally, spontaneous sputum samples were collected from staff members at Swansea University who were either current or ex-smokers, but had no known lung disease and no symptoms of COPD. All spontaneous sputum samples contained bronchial cells as confirmed by a Consultant Pathologist.

3.2.2 | Isolation of Total Genomic DNA

After collection, raw sputum samples were frozen at -80°C for up to seven days, at which time they were defrosted in ice for approximately one hour. Sputum cells were isolated as described by Lewis *et al.*, (2010) through the addition of 0.5 mL of a working solution of DTT, made up by adding 2.5 g of DTT to 31 mL of 30% aqueous methanol, and 5 mL of 30% aqueous methanol, following which they were placed on a vortex mixer for 15 minutes. Samples then underwent centrifugation at 1800 x g for 10 minutes. The supernatant was then removed and the pellet transferred to a clean 1.5 mL microcentrifuge tube and frozen at -80°C. After processing, all samples remained frozen for up to two years before total genomic DNA was extracted.

Total genomic DNA was extracted from 100 µL of isolated sputum cells within seven days of arrival to Aberystwyth laboratories using a FastDNA SPIN kit for soil (MP Biomedical, Santa Ana, USA) following manufacturer's instructions. Bead beating was carried out in a FastPrep-24 machine (MP Biomedical)

with three cycles at speed setting 6.0 for seconds, with cooling on ice for 60 seconds between cycles. Genomic DNA was eluted in to 30 µL of DES and dsDNA concentration determined using the Quant-iT dsDNA High Sensitivity assay kit and a Qubit fluorometer (Life Technologies, Paisley, UK).

3.2.3 | Metagenomic Library Preparation and Sequencing

Extracted genomic DNA was normalised to 10 ng/µL with PCR grade water (Roche Diagnostics Limited, West Sussex, UK) and 50 ng used to create metagenomic libraries using the Nextera® DNA kit (Invitrogen, San Diego, USA) following standard instructions, except that a MinElute PCR purification kit (Qiagen, Ltd Crawley, UK) was used for the clean-up of tagmented DNA. Nextera® DNA libraries were quantified as above, and approximate library sizes determined by running on a 2% agarose gel alongside HyperLadder IV (Bioline, London, UK).

Sample libraries were pooled in equimolar concentrations following Illumina guidelines and sequenced at 2 x 151 bp using an Illumina HiSeq 2500 rapid run, with samples duplicated over two lanes, and following standard manufacturer's instructions at the IBERS Aberystwyth Translational Genomics Facility.

3.2.4 | Metagenomic Sequence Analysis

After sequencing, output files for each sample were combined into one file, using the BioLinux 7 environment (Field *et al.*, 2006), for each read direction. Sequencing files were uploaded to MG-RAST (v3.2) (Meyer *et al.*, 2008) as FASTQ files, and paired-end reads joined using the facility available within MG-RAST, with non-overlapping reads retained. Sequences were dereplicated and dynamically trimmed using the default parameters for FASTQ files, and human sequences removed by screening against the *Homo sapiens* (v36) genome, available via NCBI. The MG-RAST pipeline used an automated BLASTX annotation of metagenomic sequencing reads against the SEED non-redundant database (Overbeek *et al.*, 2005).

The SEED hits can be matched to identity at various taxonomic levels; e.g. genus or species levels. Organism abundances were modelled and exported from MG-RAST using the 'Best Hit Classification' after alignment to the M5NR database, with only alignments with a maximum e-value of 1×10^{-5}, minimum identity cut-off of 97 %, and a minimum alignment cut-off of 15 being used. Functional abundances were modelled and exported from MG-RAST using 'Hierarchical Classification'. SEED matches can also be related to metabolic information, again at different levels of classification. The coarsest level of organization; the generalized cellular function was termed level 1, and the finest, individual subsystems level 3. Read abundances were transformed, to normalise for potential variations in sequencing efficacy, into percentages based upon the total abundances within each sample. Statistical analysis was completed using the MetaboAnalyst 2.0 (Xia *et al.*, 2012) facility and MINITAB 14 package. Eukaryotic taxonomic classifications were trimmed based on literature searches to remove poorly classified reads. Unless otherwise stated, where given, *P* values represent the significance of the result of one-way ANOVA tests. All sequences files are publicly available through the MG-RAST IDs detailed in Chapter 3 Appendix, Supplementary Table 3.1. Raw sequence files have been deposited at the European Nucleotide Archive under primary accession number PRJEB9034 and secondary accession number ERP010088.

3.3 | Results

Spontaneous sputum was collected from eight patients (five male: three female) with a clinical and spirometric diagnosis of COPD from two UK hospitals, with a mean age of 46, mean smoking pack years of 46 and mean FEV_1% of predicted of 46%. Of the eight COPD patients, three were classified as GOLD stage II and five as GOLD stage III. Ten (six male: four female) spontaneous sputum samples were collected from staff members, (mean age = 53) at Swansea University who were either current or ex-smokers but had no known lung disease and no symptoms of COPD. Individual participant details, alongside additional clinical information for COPD patients, are given in Table 3.2. Full participant information is detailed in Chapter 3 Appendix, Supplementary Table 3.1. No discernible differences were observed between the characteristics detailed in Table 3.1, in regards to age, male to female ratio and smoking status.

TABLE 3.2 | Characteristics of Participant/Patients for each Disease Group
Data are means for each disease group, with standard deviations displayed in brackets. FEV_1% = forced expulsion volume in one second, as a percentage of the predicted value. Smoking pack years is individual history normalised to 20 cigarettes a day for one year. GOLD = Global Initiative for Chronic Obstructive Pulmonary Disease rating of COPD severity. nc = not collected.

Group	Control	COPD
Number	10	8
Age	52.90 (*13.26*)	67.75 (*5.26*)
Male : Female Ratio	6 : 4	5 : 3
Smoking Status		
Current	4	5
Ex	6	2
Never	0	1
Smoking Pack Years	nc	45.57 (*12.87*)
FEV_1% of Predicted	nc	45.88 (*11.47*)
GOLD Rating		
I	nc	0
II	nc	3
III	nc	5
IV	nc	0
Antibiotic Use	0	2

3.3.1 | Preliminary Sequence Read Analysis

Genomic DNA sequencing statistics are shown in Chapter 3 Appendix, Supplementary Table 3.2. One-way ANOVA showed no statistically significant differences in all but one sequencing output. Average read lengths were significantly longer (P = 0.001) in control samples, by approximately 4 bp. However, because no other read characteristics were shown to be statistically significant, it is likely that sequencing with the HiSeq 2500 platform and subsequent metagenomic analysis with the MG-RAST pipeline did not introduce any degree of bias that would affect subsequent data analysis. All 18 sequenced metagenomes are publicly available via the links detailed in Chapter 3 Appendix, Supplementary Table 3.1.

3.3.2 | Comparison of the Taxonomy of COPD Microbiome

MG-RAST Principal Component Analysis revealed separation between the control and COPD groups when considering taxonomic classification; with five out of the eight COPD samples forming a distinctive cluster, Figure 3.3a. However, three COPD samples remained clustered away from the main COPD group

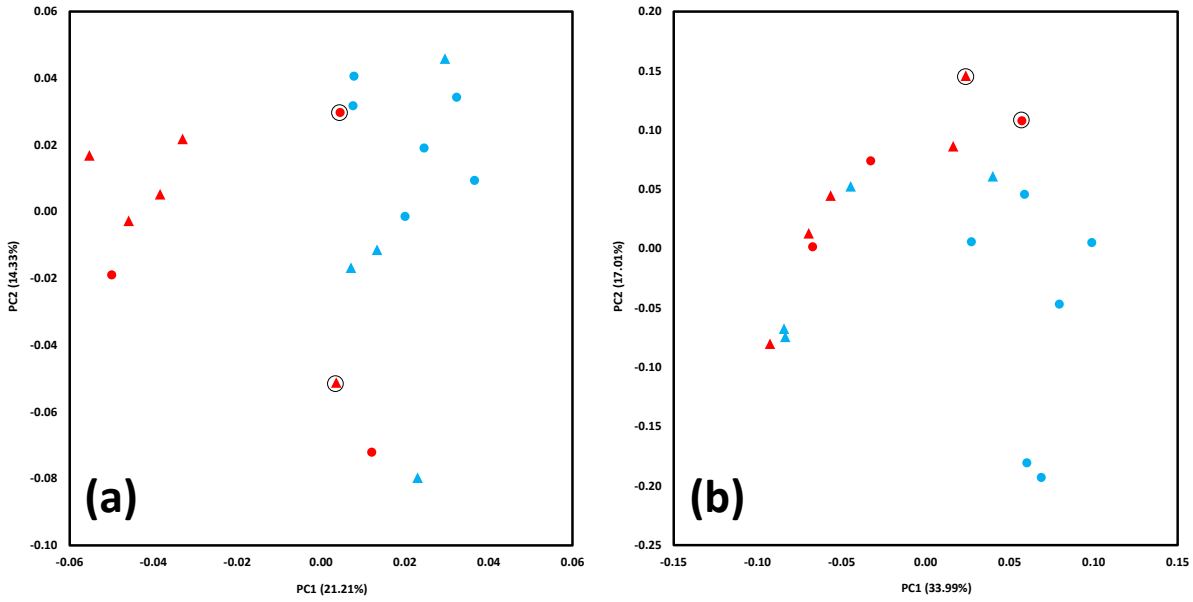


FIGURE 3.3| Principal Component Analysis of Taxonomy and Functional Classifications
Using MG-RAST analysis platform, PCA plots were created using (a) taxonomic and (b) functional classifications, using the analysis method detailed previously. Control samples are coloured blue and COPD red. Triangles indicate patients who are current smokers, and black circles indicate the patient has antibiotic use in their medical history prior to giving a sample. PCA plots drawn using normalised values and Manhattan distance.

with the Control group. There were no unifying characteristics of these three samples, with two being GOLD stage III and one GOLD stage II. PCA separation did not appear to be influenced by smoking status or reported prior use of antibiotics.

A total of eight bacterial genera were present in all 18 samples, *Haemophilus, Lactobacillus, Neisseria, Ochrobactrum, Pseudomonas, Staphylococcus, Streptococcus,* and *Veillonella*. At the species level, Figure 3.4, there are four species present in all 18 samples, *Haemophilus influenzae*, *Ochrobactrum anthropic*, *Streptococcus pneumoniae*, and *Streptococcus thermophilus*. Crucially, four additional species found in all of the COPD samples but not all control samples: *Staphylococcus aureus*, *Stenotrophomonas maltophilia, Streptococcus agalactiae*, and *Streptococcus pyogenes*. Significantly, all of these species are human pathogens, with *S. pyogenes* being the most pathogenic Streptococcal bacterium. However, no

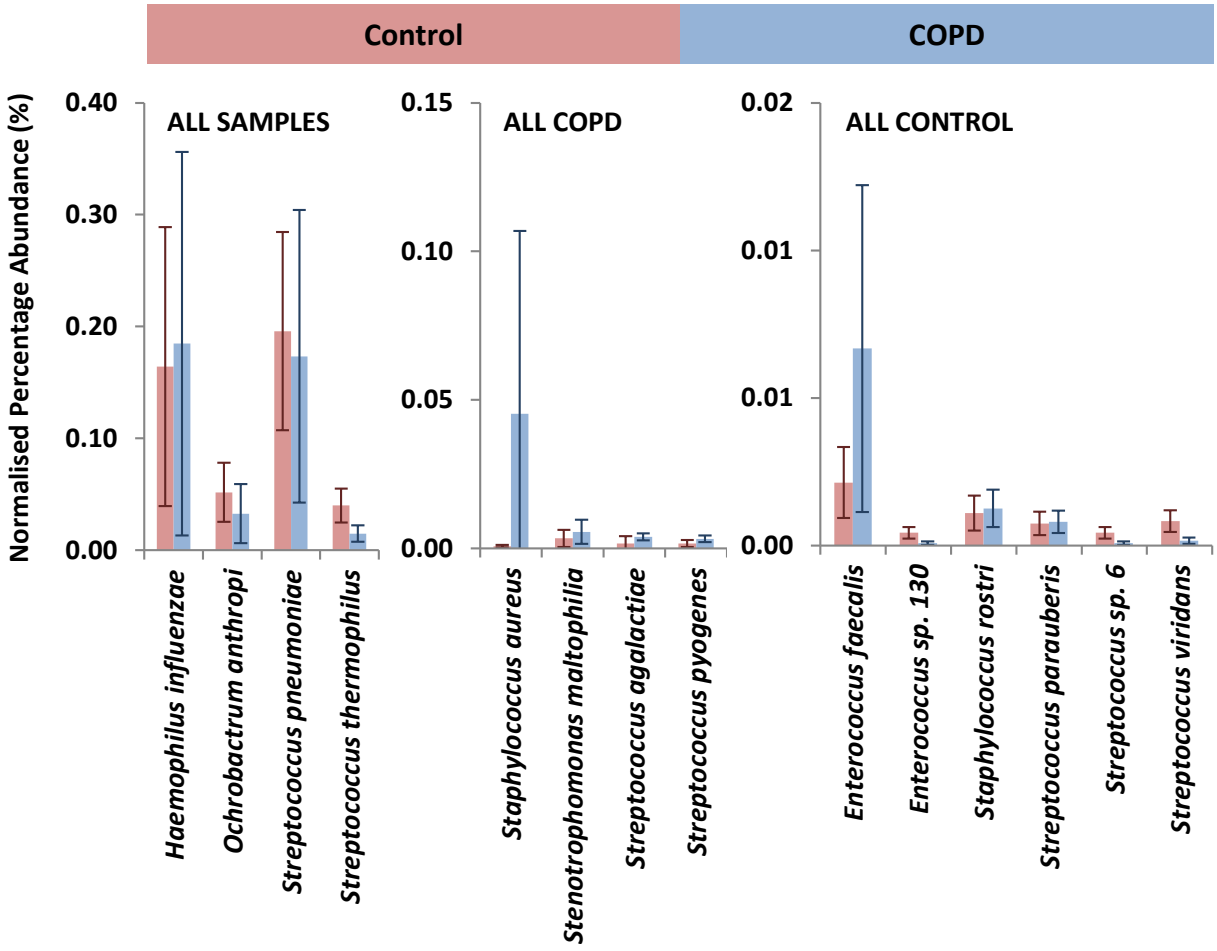


FIGURE 3.4 | Core Microbiome of All Samples, COPD Samples, and Control Samples
Abundance of the 14 bacterial species that constitute the 'core microbiome' in Control participants and COPD patients. Four bacterial species were found in all samples from both groups, four species were found in all of the COPD samples but not all of the Control samples, and six species were found in all of the Control samples but not all of the COPD samples. There was no bacterial species that was common to all samples in one of the two groups, but unique to that group.

statistically significant differences, between COPD and Control groups, were evident for the four bacterial species found in all COPD samples, but not all Control samples, suggesting that individual differences may play an important role in determining the level of bacterial species in the 'core' microbiome. Additionally, six bacterial species were found in all control samples but not in all COPD samples; two *Enterococcus* species, *S. rostri* and the *Streptococcus* species *S. parauaberis, S. virdans* and sp.6. Some non-human eukaryotic sequences were identified in metagenomic libraries, but no significant differences in species abundance were detected when comparing control and COPD groups. Other species which exhibited statistically significant fold changes in abundance between COPD and controls were targeted, Figure 3.5. These species included the pathogens *Gemella haemolyses*, *Abiotrophia para-adiacens* and *Glemella sanginis.* The *Streptococcus* genus in particular appeared to exhibit many changes in abundance in COPD patient samples, with both increases and decreases compared to the Control group.

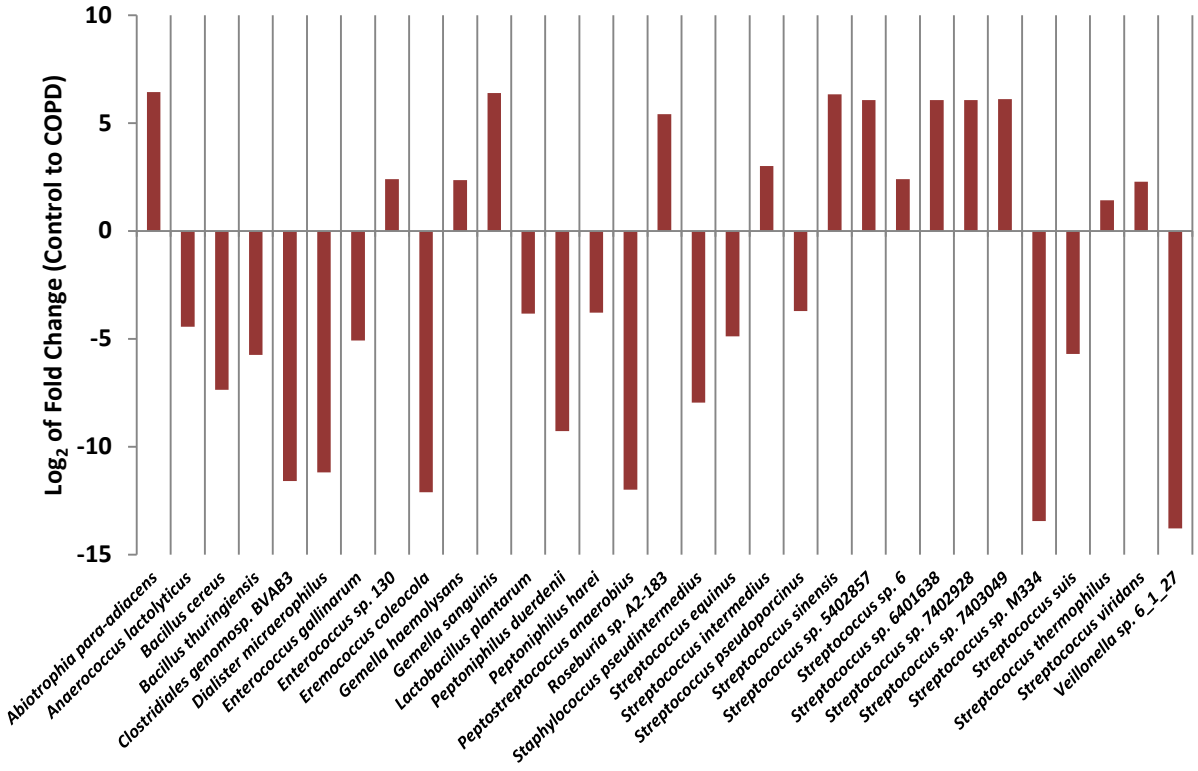


FIGURE 3.5 | Significant Changes in Species Abundance from Control to COPD
Using MetaboAnalyst 2.0, t-Tests and fold-changes were calculated from normalised percentages of reads, with only those with a *P* value of < 0.05 charted. The *Streptococcus* genus appears to be highly dynamic in COPD, with both increases and decreases in species abundance in COPD compared to the Control group. On average, COPD samples appear to show a net loss in the abundance of obligate anaerobes, but a net increase in the abundance of facultative anaerobes.

3.3.3 | Comparison of Functional Capacity of COPD Microbiome

Functional classification, Figure 3.3b, appeared to reduce the separation of the two groups, when compared to taxonomic classification, Figure 3.3a. However, seven of the control samples clustered away from the COPD samples, but three remained clustered with the majority of the COPD samples. As with taxonomic classifications, smoking status did not appear to influence PCA separation, but the presence of antibiotic use in medical history may be a more significant factor in functional separation than for taxonomic separation.

Significant fold changes in gene functional classifications were also considered. At the crudest functional classification; Level 1, there were significantly fewer alignments to genes involved in carbohydrate metabolism in COPD patients, but an increase in clustering-based subsystems, horizontal gene transfer, and nucleosides and nucleotides. At the more resolved Level 2 functional classification, 26 classifications exhibited a significant difference, with 22 higher in COPD patients. At the most resolved Level 3, only significant increases in COPD patients were observed, Figure 3.6. These alignments appear to centre on functional classifications involved in bacterial growth, including bacterial cell division, nucleosides and nucleotides, and amino acid, protein and RNA metabolism. Additionally, functional classifications involved in the stress response associated with the heat shock DnaK gene cluster were significantly higher in COPD samples than Control samples.

3.3.4 | Microbiome Changes Associated with COPD Severity

In assessing the potential influence of our finding on the severity of airflow obstruction (FEV_1 % of predicted), Table 3.3, a positive correlation with the *Streptococcus* genus (R^2 = 51.8%, P = 0.044), and more specifically *S. pneumoniae* (R^2 = 63.6%, P = 0.018) was found. Additionally, functional positive correlations were observed with the classification of di- and oligosaccharides (R^2 = 50.8%, P = 0.047) at Level 2, and more specifically with sialic acid metabolism (functional level 3) (R^2 = 51.1%, P = 0.046) at

Level 3. No significant correlation between *S. pneumonia* and smoking pack years or age was observed, suggesting that COPD severity is the main factor influencing the observed correlation.

The genus *Neisseria* showed a correlation with smoking pack years (R^2 = 66.1%, *P* = 0.014). Notable significant positive functional correlations for smoking pack years were also observed with bacterial DNA repair, potassium homeostasis and the protease modulator YbbK. With regards to age, the *Ochrobactrum* genus showed a significant positive relationship (R^2 = 51.6%, *P* = 0.045), and specifically *O. anthropi* (R^2 = 51.6%, *P* = 0.045). There were also significant correlations with the associated biochemical pathways linked to glutamate and proline metabolism, and separately with quorum sensing and biofilm formation which could be associated with monosaccharide production.

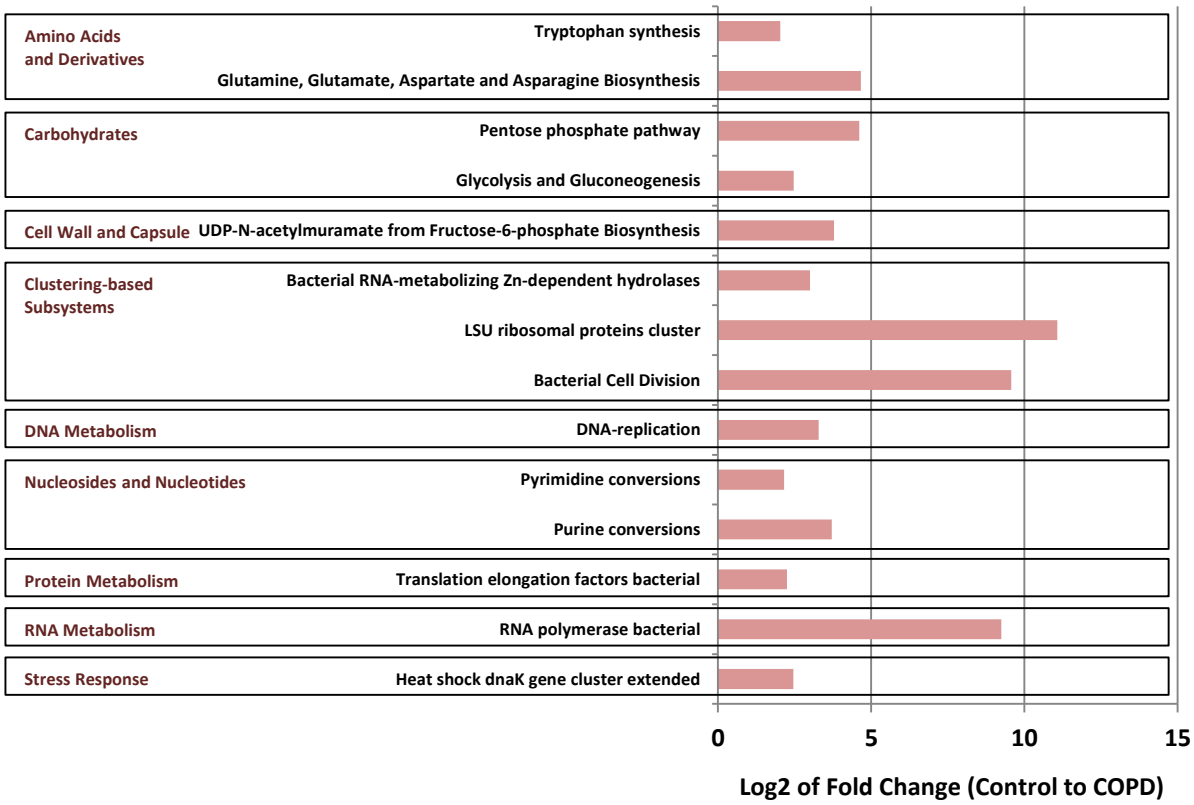


FIGURE 3.6 | Significant Changes in Functional Classification Abundance from Control to COPD
Using MetaboAnalyst 2.0, t-Tests and fold-changes were calculated from normalised percentages of reads, with only those with a *P* value of < 0.05 charted. Differences at Level 3 appear to centre on changes to those reads aligned to those with functional roles in bacterial cell division.

TABLE 3.3 | Regression Analysis for COPD Patients using FEV_1%, Smoking Pack Years and Age

Regression analysis of FEV1% of predicted, commonly used as a measure of COPD severity, with normalised sequence numbers, reveals that the Streptococcus genus, and specifically S. pneumonia is positively correlated with FEV1% of predicted, but not smoking pack years or age, suggesting that S. pneumonia could act as a biomarker for COPD disease progression. Only those taxonomic or functional classifications found in all eight COPD patients, and with an R^2 value of >0.5 were used in regression analysis. + or – symbols indicate whether relationship is positive or negative respectively. Significant regressions are highlighted in bold. *P* represents *P* value of significance of correlation seen.

				FEV_1% of Predicted			Smoking Pack Years			Age		
			+ / -	R^2	R^2 - adj	*P*	R^2	R^2 - adj	*P*	R^2	R^2 - adj	*P*
Taxonomy	**Genus**	***Streptococcus***	+	**51.8**	**43.7**	**0.044**	23.1	10.3	0.228	20.6	7.3	0.259
		Neisseria	-	2.4	0.0	0.715	**66.1**	**60.4**	**0.014**	6.3	0.0	0.548
		Ochrobactrum	-	13.4	0.0	0.373	3.1	0.0	0.679	**51.6**	**43.6**	**0.045**
	Species	***Streptococcus pneumonia***	+	**63.6**	**57.6**	**0.018**	8.4	0.0	0.487	40.1	30.1	0.092
		Ochrobactrum anthropi	-	13.4	0.0	0.373	3.1	0.0	0.679	**51.6**	**43.5**	**0.045**
Function	**Level 2**	**Di- and oligosaccharides**	+	**50.8**	**42.6**	**0.047**	1.8	0.0	0.752	29.7	18.0	0.162
		Glutamine, glutamate, aspartate	-	47.5	38.7	0.059	8.2	0.0	0.491	**78.7**	**75.2**	**0.003**
		Monosaccharides	+	39.2	29.1	0.097	**3.8**	**0.0**	**0.000**	**55.2**	**47.7**	**0.035**
		Quorum sensing and biofilm form	-	19.2	5.7	0.278	0.2	0.0	0.912	**52.9**	**45.0**	**0.041**
	Level 3	**Sialic Acid Metabolism**	+	**51.1**	**42.9**	**0.046**	18.1	4.5	0.293	40.0	30.0	0.092
		DNA repair, bacterial	-	2.4	0.0	0.714	**74.8**	**70.6**	**0.006**	6.5	0.0	0.543
		Potassium homeostasis	-	15.0	0.8	0.344	**65.6**	**59.8**	**0.015**	21.1	7.9	0.252
		YbbK	-	3.4	0.0	0.662	**61.0**	**54.5**	**0.022**	18.0	4.4	0.294
		Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis	-	30.3	18.7	0.157	10.3	0.0	0.439	**66.5**	**60.9**	**0.014**
		Proline, 4-hydroxyproline uptake	-	17.1	3.3	0.309	15.4	1.4	0.335	**53.3**	**45.5**	**0.040**

3.4 | Discussion

The role of microbial pathogens in COPD has been well documented, specifically in relation to exacerbations (Patel, 2002). However, although there have been studies of microbial changes in patients with differing levels of COPD severity (Erb-Downward *et al.*, 2011; Pragman *et al.*, 2012; Sze *et al.*, 2012), these have not unambiguously identified the species present, or suggested changes in the functions of the bacterial populations. In this portion of work, the first metagenomic study of the lung microbiome in patients with COPD, which reveals both its function and structure, down to the species level of resolution, was detailed.

In the human gut, metagenomic sequencing has been used to develop novel hypotheses into personalised disease risk factors based on gene numbers in the microbiome (Qin *et al.*, 2010), and metagenomics has yielded novel insights into the lung microbiome of cystic fibrosis (Lim *et al.*, 2014). As DNA sequencing becomes more accessible, metagenomic approaches will be increasingly applied in personalised medicine, which could provide avenues for improved methods of diagnosing, monitoring, and treating COPD. Any personalised medicine strategy however, needs a minimally invasive sampling technique, and thus, spontaneous sputum was used in this work, rather than the more invasive method of BAL sampling (Parr *et al.*, 2006).

3.4.1 | Species Composition of the COPD Lung Microbiome

Despite considerable variation between individual samples, significant differences between the COPD and control groups were found. These are mostly likely to reflect shifts in the species makeup of the lung microbiome, so that they become sufficiently prominent to be detected using metagenomic sequencing technology. One of the key aspects of these differences was the noteworthy increases in four bacterial species - all pathogens - to above detection limits only in COPD patients. Interestingly, none of the eight COPD patients were exacerbating, but *S. aureus* and *S. maltophilia*, which are commonly isolated from COPD patients with acute exacerbation, were found (Nseir *et al.*, 2006).

Furthermore, *S. maltophilia* has also been linked to exacerbations in cystic fibrosis patients, which suggests that it may be an important opportunistic lung pathogen in many morbidities (Ciofu, Hansen and Høiby, 2013). Most importantly, besides offering increased understanding of the developing underlying pathology, these four bacterial species could act as biomarkers for COPD, which could be monitored, in addition to FEV_1% of predicted (Agusti and MacNee, 2013).

A large number of bacterial species were shown to be significantly different in Control and COPD samples. Interestingly, a higher number of bacterial species show a significantly lower abundance in COPD samples, than those that show a significantly higher level. The *Streptococcus* genus in particular showed a high degree of dynamism. The evident flux in the lung microbiome present in COPD patients suggests that the environment of the COPD lung poses a selective pressure on the microbiome. This may mean that only bacteria with the ability to tolerate the changing environment, such as a reduction in the oxygen level, will survive, explaining the changes seen in this portion of work.

Currently, COPD diagnosis relies heavily upon the use of the FEV_1/FVC ratio to define airflow limitation. However, this has been shown to result in higher levels of COPD diagnosis in elderly patients, and possibly under-diagnosis in younger patients, below 45 years, who may have milder disease (Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014). Therefore, additional methods of diagnosing COPD, alongside spirometric analysis of FEV_1/FVC, may improve COPD diagnosis, and may even be able to act as a biomarker before the disease has developed to a stage where it is measurable through spirometry.

3.4.2 | Functional Characteristics of the COPD Lung Microbiome

Considering metagenomic revealed changes in function, there were increases in the abundance of functional alignments associated with bacterial growth, particularly bacterial cell division, nucleosides and nucleotides, and amino acid, carbohydrate, DNA, protein, and RNA metabolism. With increases also in factors linked to horizontal gene transfer, this indicates large-scale genetic exchanges and the

capacity for rapid bacterial growth may be a characteristic of the COPD microbiome. Similar bacterial genomic flux also appears to be a feature of cystic fibrosis patients (Bittar and Rolain, 2010), and bacterial load has been linked to periods of COPD exacerbation (Garcha *et al.*, 2012).

Acute exacerbations of COPD can be relatively sudden, and can be precipitated by a number of factors, including colonisation by virus, bacteria, or a combination of both. A potential hypothesis of bacterial-associated exacerbations, which has emerged from this portion of work, is that the microbiome of COPD patients has the functional capacity for rapid bacterial growth, which may explain the sudden onset of exacerbations seen. Functional alignments shown to have a significantly higher level in the COPD microbiome include those involved in DNA replication, protein metabolism, bacterial RNA polymerase, and bacterial cell division, amongst others, which are essential factors in bacterial cell growth and division. As discussed previously, the COPD lung may provide a selective pressure on the lung microbiome, leading to alterations in its functional capability. Therefore, analysis of the functional capacity of the lung microbiome in COPD patients may provide for an approach to personalised medicine that has the potential to create risk factors for acute exacerbations of COPD.

Functional analysis also revealed the significant increase in alignments to the heat shock DnaK gene cluster, which in bacteria, is responsible for producing the heat shock protein Hsp70. Analogues of Hsp70 have been shown to have significant anti-inflammatory responses in many inflammatory diseases (Borges *et al.*, 2012), and could provide a mechanism for bacterial defence from the inflammatory mediators inherent in the environment of the COPD lung.

3.4.3 | Lung Microbiome Characteristics Associated with COPD Severity

Previous COPD lung microbiome studies have reported patients with severe COPD had a high prevalence of *P. aeruginosa, H. influenzae* and *S. pneumoniae* (Monso *et al.*, 2003; Groenewegen and Wouters, 2003). Analysis of this portion of work failed to suggest any significant correlation between *P. aeruginosa* and FEV_1% of predicted. However, as *P. aeruginosa* does not appear to be part of the

detected core lung microbiome of our baseline COPD patients, it may be that any change in the abundance of this opportunistic pathogen is linked to exacerbation rather than COPD progression as indicated by measures of FEV_1% of predicted (Renom *et al.*, 2010). Conversely, *H. influenzae* was part of the core microbiome but was not changed in abundance in these COPD patients. Taken together, these observations indicate that abundance changes in these bacteria species would be poor biomarkers for COPD progression. A significant positive correlation, however, was observed with *S. pneumonia* and FEV_1% of predicted, suggesting that as airflow obstruction increases, the percentage abundance of *S. pneumonia* decreases in the stable state. *S. pneumonia* is frequently cultured from the sputum samples of patients during exacerbations (Monso *et al.*, 2003; Groenewegen and Wouters, 2003), and it is the main target of initial treatment with penicillins. Detecting subtle changes in *S. pneumonia* loads, may allow prediction of COPD progression and allow earlier interventions.

A further, significant, positive correlation was observed between FEV_1% predicted, and the percentage abundance of genes associated with sialic acid metabolism. Sialic acids are nine carbon sugars backbone monosaccharides mainly decorating the outside of vertebrate cells, but also some microbes (Severi, Hood and Thomas, 2007; Varki, 2007; Vimr and Lichtensteiger, 2002). Extracellular sialic acid moieties have many roles in vertebrate immunology and can act to mask cell surface receptors; or act as recognition sites for various lectins and antibodies. These roles including the modulation of leukocyte trafficking via selectins and influencing complement activation (Varki and Gagneux, 2012).

Sialic acid binding immunoglobulins (Ig)-like lectins (siglecs) are found in immune cells and will recognise different linkage-specific sialic acids. Examples of siglecs are siglec-3/CD33 related-siglecs found on haematopoietic cell lineages, siglec-9 on natural killer (NK) cells and siglec-8 only on circulating eosinophils. After binding sialyated moieties, siglecs can drive the internalisation of sialyated pathogens and crucially, modulate pathogen-/damage-associated molecular patterns-mediated (PAMP/DAMP) inflammation along with inhibition of NK cell activation. Sialic acid-Siglec interaction therefore serve to

maintain a baseline non-activated state of innate immune cells, and limit inflammatory responses activation through PAMP/DAMP recognition (Cao and Crocker, 2011).

The pathological advantages to the pathogen of acquired sialic acid decoration is therefore to augment siglec mediated avoidance of PAMP/DAMP recognition (Cao and Crocker, 2011; Chang and Nizet, 2014; Carlin *et al.*, 2007). Additionally, the presence of sialylated lipopolysaccharide on the bacterial surface can prevent complement activation by binding to the C3 component of the complement cascade (Shaughnessy *et al.*, 2009). This portion of work has shown a significant correlation with reduction of bacterial sialic acid metabolism with decreasing FEV_1 % of predicted scores, which would indicate a shift towards a lesser capacity to avoid recognition and thus suppress inflammation – a key feature of COPD. Thus, a reduction in sialic acid metabolising bacteria could be an important pathological feature in COPD progression.

Sialic acid is also an important carbon source in the respiratory tract, particularly the lungs, for bacterial colonisation. Indeed, the bacterial pathogen *S. pneumoniae*, which has been shown in this portion of work to be closely associated with COPD severity, is able to utilise human sialic acid as a carbon source for growth. Additionally, its ability to metabolise sialic acid also carries the advantages of being able to compete with other bacteria for an environmental niche, aiding the bacteria's progression through the mucin layer, and promoting adherence to epithelial cells (Marion *et al.*, 2011). This portion of work has demonstrated a significant correlation between COPD severity and both *S. pneumoniae* and functional alignments involved with sialic acid metabolism. This raises the interesting possibility of the interaction between these correlations, either as being causative or reflective of changes in the COPD microbiome, and warrant significant further study.

As previously discussed, viral colonisation can be an important event in the onset of acute exacerbations of COPD. Many viral infections are initiated through binding to host cell surface receptors, including those that contain sialic acid (Stencel-Baerenwald *et al.*, 2014). In this portion of work, sialic acid has

been shown to have a significant positive correlation with FEV_1% of predicted, and thus COPD severity. Although viruses were not looked at in this work, they are nonetheless an important constituent of the lung microbiome, particularly in COPD. It may be that the microbiome-associated alterations of the sialic acid present in the lung microbiome of COPD will alter the ability of viral pathogens to colonise, thus impacting on the risk of acute exacerbations of COPD. Therefore, the measurement of sialic acid within the lungs of COPD patients may be able to act as a biomarker for risk of acute exacerbations of COPD. Colorimetric methods for measuring sialic acid in both serum and BAL fluid already exist, and may provide a potential method of non-invasive risk profiling of acute exacerbations in COPD patients (Isitmangil *et al.*, 2001).

The observed significant correlations with FEV_1% of predicted were substantiated as similar relationships with either smoking pack years, or the age of COPD patients, were not evident. This reinforces the concept that these observed changes are as a result of worsening airflow obstruction, rather than a direct modulating effect of smoking or age on lung immune function. Nevertheless, this portion of work does indirectly identify significant relationships between smoking and the *Neisseria* genus, which has previously been linked to smoking (Morris *et al.*, 2013), and a number of Level 3 functional classifications, namely bacterial DNA repair and potassium homeostasis. These features possibly reflect smoking linked bacterial genomic damage and a response to the inclusion of potassium salts in cigarette papers, respectively (Husgafvel-Pursiainen, 2004; Zawadzki *et al.*, 2005).

Despite its small size, this portion of work describes the first use of metagenomic sequencing to reveal novel insights into the microbiome present in the COPD lung. Additionally, potential novel bacterial and functional biomarkers for COPD progression were identified. This demonstrated the potential strengths of using metagenomic techniques to characterise the upper respiratory microbiome in patients with COPD.

3.5 | Conclusions and Future Work

This portion of work aimed to take advantage of metagenomic sequencing technologies to investigate the microbiome of patients with COPD to the species-level of taxonomy, and for the first time, reveal novel insights into the functional capabilities of the microbiome. To this end, the species-level composition of the lung microbiome in COPD was evaluated, and a core microbiome revealed; containing bacterial species that could be used as biomarkers for COPD. Additionally, the significant changes observed in the functional capacity of the lung microbiome in COPD may be help to explain the role of the microbiome in periods of acute exacerbations. Furthermore, the lung microbiome in COPD patients could be used to develop risk profiles of patients in regards to their susceptibility for acute COPD exacerbations.

One of the most significant portions of this work was the identification of *S. pneumoniae* and functional alignments to sialic acid metabolism as being positively correlated with FEV_1% of predicted, and thus COPD severity. The identification of sialic acid specifically may allow for the development of novel diagnostic and risk susceptibility techniques, and also suggest targets for therapeutic interventions.

Nevertheless, a number of important questions still remain in regards to the role of the lung microbiome in COPD patients. Future work to address these may include the longitudinal monitoring of COPD patients to sample the lung microbiome before, during, and after periods of acute exacerbations to gauge whether changes in the microbiome are predictive, reflective, or even causative. Additionally, mechanistic studies of the role of sialic acid in the lung microbiome, particularly in relation to colonisation by microbial pathogens important in COPD, such as *S. pneumoniae* and *H. influenza*, may help to explain the role that these elements have in COPD progression, and potentially acute exacerbations.

CHAPTER 4 | Defining the Temporal Variability of the Salivary Microbiome and Metabolome

CHAPTER SUMMARY | The human microbiome and metabolome are both important components of homeostasis, and in the monitoring of health and disease. It is well established that incidence of upper respiratory tract infections vary throughout the year, and it may be expected that the salivary microbiome would reflect this. To determine this, the saliva of 40 healthy participants was collected every two months over a 12 day period, from October 2012 to October 2013, alongside lifestyle information. Samples were analysed in terms of bacterial load, through quantitative PCR, measurement of pH, and metabolomic fingerprinting using negative mode LTQ-MS. A sub-group of ten participants, selected because of their similar lifestyle information, underwent 16S rRNA (V3 to V4) amplicon sequencing. Subsequent sequences were demultiplexed, merged, trimmed for quality, and uploaded to the MG-RAST online pipeline for taxonomic assignment. Estimated bacterial load was shown to be significantly ($P < 0.001$) higher in February 2013 than at all other sampling time points, with individuals' changes between time points displaying significant ($P = 0.003$) flux. Salivary pH levels were shown to be significantly ($P < 0.001$) higher in December 2012 than in October 2012 and February 2013, with significant ($P < 0.001$) individual variations seen across the sampling period. In regards to the stability of the taxonomic composition of the salivary microbiome, α-diversity values showed significant differences between participants ($P < 0.001$), but not between sampling periods ($P = 0.801$), and a small, but significant positive correlation with salivary pH ($R^2 = 7.8\%$; $P = 0.019$). At the phylum level of classification, significant differences were evident between participants in the Actinobacteria ($P < 0.001$), Bacteroidetes ($P < 0.001$), Firmicutes ($P = 0.008$), Fusobacteria ($P < 0.001$), Proteobacteria ($P < 0.001$), Synergistetes ($P < 0.001$), and Spirochaetes ($P = 0.003$) phyla. The salivary metabolome also displayed temporal stability, with no separation between participants through principal component analysis. The salivary microbiome shows temporal stability in terms of bacterial taxonomic diversity, but not load, over a one year period, with individual differences likely to be the main factor in determining its composition. This work suggests that comparing the human salivary microbiome and metabolome at different time points over the year is valid, as minimal temporal variability is shown.

4.1 | Introduction

Investigations of the human microbiome and metabolome have allowed for the importance of both in health and disease to be realised. The gut microbiome, for example, has been shown to be an important component of gastrointestinal diseases such as Crohn's (Manichanh *et al.*, 2006). Additionally, the human metabolome is an important source of biomarkers for disease (Spratlin, Serkova and Eckhardt, 2009). However, many of these findings, particularly metabolome-derived biomarkers, are dependent upon the principle that the human microbiome and metabolome are temporally stable. To date, this is a principle that has not been firmly established.

4.1.1 | The Oral Cavity and Saliva Production

The human oral cavity is a major gateway to the human body, and serves many purposes in both health and disease; acting as the main point of entry to both the digestive and respiratory tracts. The oral cavity has a number of distinct physical features including the hard and soft palates, the tonsils, uvula, tongue, vestibule, gingiva and teeth, Figure 4.1a.

One of the spatial components of the oral cavity, albeit without a fixed position is saliva. Healthy human adults produce between 0.5 to 1.5 L of saliva a day, at a rate of 0.5 mL per minute, though a number of

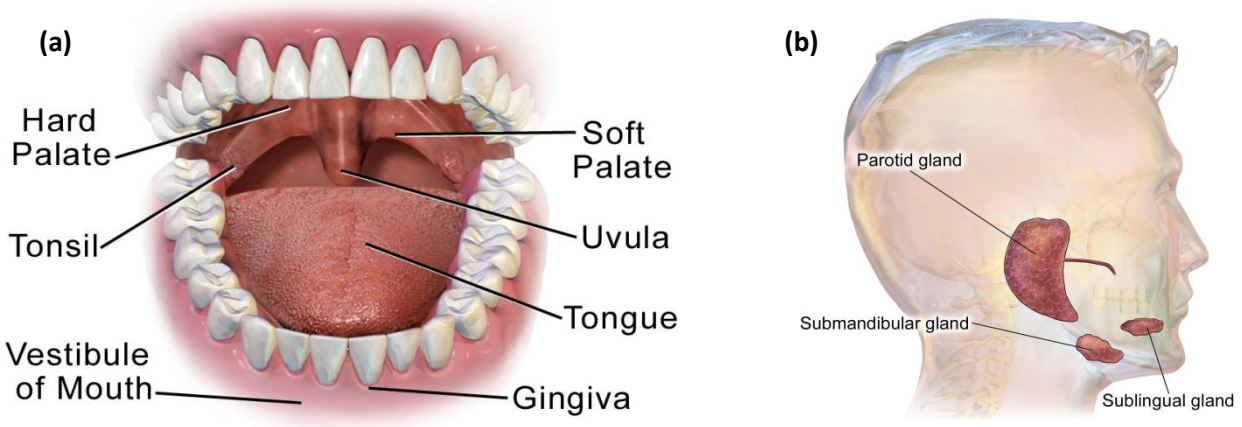


FIGURE 4.1 | The Structure of the Human Oral Cavity
The (a) human oral cavity consists of a number of spatially distinct structures, including the tongue, soft and hard palates, the gingiva, tongue, tonsils, uvula, and vestibule. Within the human oral cavity, saliva is produced by the (b) three salivary glands; the parotid gland, the submandibular gland, and the sublingual gland. Figures are copyright free and adapted from Wikipedia.com.

factors can impair production in terms of both volume and composition (Pfaffe *et al.*, 2011). The majority of human saliva is produced by three major salivary glands, Figure 4.1b, complemented by minor glands which produce saliva to coat the mouth surface. Many biomolecules found within human saliva have their origins in the circulatory system and thus, saliva offers a non-invasive diagnostic tool for a number of diseases (Matias *et al.*, 2012). Saliva is predominately water, with only a small amount of its volume consisting of biomolecules such as proteins, peptides, nucleic acids, electrolytes, hormones, and enzymes. These serves as important modulators in the maintenance of homeostasis within the oral cavity, helping to maintain a stable pH and prevent dysbiosis in the microbiome (Pfaffe *et al.*, 2011).

4.1.2 | The Oral Microbiome

The human oral cavity consists of a number of defined spatial regions, Figure 4.1a, which have been shown to have distinct microbiomes (Aas *et al.*, 2005). A large number of microbial taxa, approximately 600 at the species level, are present within the human oral cavity, many of which are not associated with disease initiation, aetiology, or pathogenesis (Dewhirst *et al.*, 2010). Nevertheless, the host microbiome has been shown to be an important part of oral diseases, such as periodontal disease and dental caries.

Periodontal disease, for example, affects tooth-supporting structures and can also exacerbate existing comorbidities, such as cardiovascular and pulmonary diseases. Analysis of the human oral microbiome in patients with and without periodontal disease has revealed a number of metabolic pathways enriched in the microbiome which are associated with virulence factors, and a shift from a Gram-positive to a Gram-negative taxonomic domination. This suggests the oral microbiome in periodontal disease is an important component of the disease, constituents of which are able to exploit host dysbiosis which precedes the full clinical manifestation of the disease (Liu *et al.*, 2012). Similarly, dental caries has been linked to a shift in the taxonomic composition of the oral microbiome, rather than the emergence of a distinct bacterial species (Yang *et al.*, 2012).

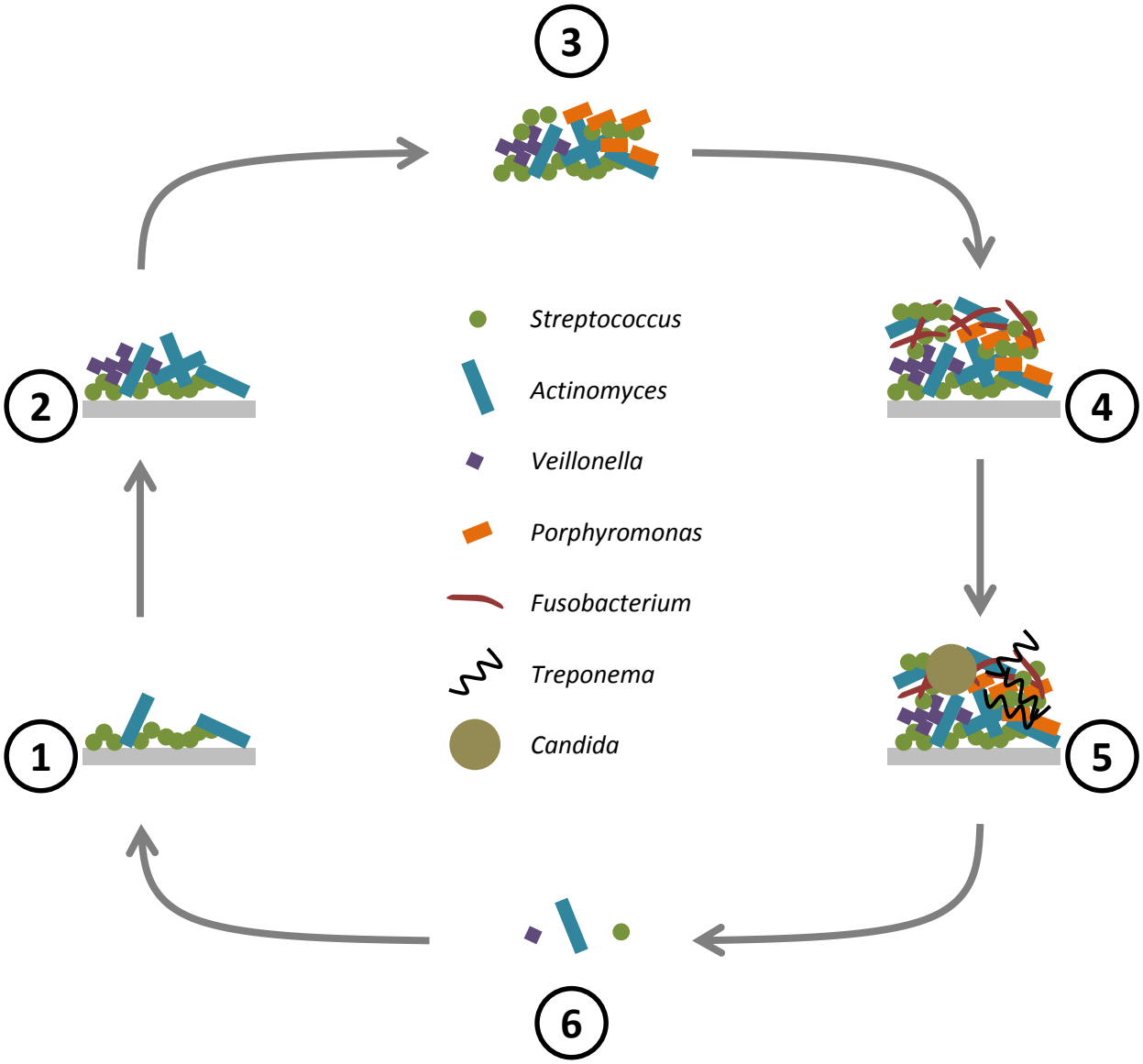


FIGURE 4.2 | Biofilm Formation in the Oral Cavity
Biofilm formation in the oral cavity follows a cyclical pattern whereby initial colonising bacteria provide a scaffold for the attachment of other members of the oral microbiome. On a clean tooth surface, colonisation will begin with members of the *Streptococcus* and *Actinomyces* genera. Additional microbes, such as *Veillonella* and *Porphyromonas* genera follow, and because of changes in the conditions of the environmental niche, incorporation of anaerobic bacteria such as the *Fusobacterium* genus is encouraged. This creates conditions that allow for incorporation of potential oral pathogens such as members of the *Treponema* and *Candida* genus. Figure adapted from Jenkinson and Lamont, (2005).

Advances in understanding periodontal disease and dental caries by approaching the conditions in terms of whole-scale changes in the microbiome, rather than a traditional focus on the responsibility of single bacterial species have refocused the field of oral microbiology. The role of single species is nevertheless important in many oral diseases, such as the mutans group Streptococci in dental caries. However,

because approximately 50% of oral bacteria are yet to be cultured, the importance of these cultureable aetiological agents may be overestimated, and in fact may be dependent upon the actions of a number of yet to be cultured microbes (Jenkinson and Lamont, 2005).

The formation of biofilms within the oral cavity, particularly on the tooth surface, exemplifies the interdependence of single bacterial species with the oral microbiome as a whole. As Figure 4.2 illustrates, the formation of biofilms follows a cycle whereby initial colonising bacteria act as scaffolds for the attachment of other members of the oral microbiome. Initial colonising bacteria, including members of the *Streptococcus* and *Actinomyces* genera, are able to do so because they are able to bind to salivary receptors that coat uncolonised tooth surfaces (Jenkinson and Lamont, 2005). This is particularly evident for *Streptococcus* bacteria which have been shown to bind to salivary proteins including, amongst others, α-amylase, fibronectin, and lactoferrin (Scannapieco, 1994). This cycle of colonisation observed on the tooth, which culminates in the addition of *Treponema* and *Candida* species, shows the importance of the microbiome in oral diseases such as dental caries and periodontal disease, for which members of these genera are believed to be causative organisms (Jenkinson and Lamont, 2005). Without the proceeding steps in biofilm formation on the tooth, neither *Treponema* nor *Candida* would be able to attach to the tooth surfaces, and thus would not be able play their role in a number of oral diseases.

The role of the wider microbiome in the onset of oral disease, and the interconnectedness of species within the microbiome in both health and disease, is not solely due to the creation of a scaffold that allows for adhesion of secondary colonisers. Constituents of the oral microbiome are also interdependent at a metabolic level. Within the human mouth, one of the main sources of nutrition for the microbiome is saliva, but its constituents are not readily accessible to bacteria. As a result of the diverse range of enzymes required to degrade salivary proteins and glycoproteins, it is unlikely that a sole bacterium would possess these in isolation. Co-culture studies have shown that when ten oral bacterial species are grown in an artificial salivary medium, the total accumulated biomass is

significantly greater than when five bacterial species are grown (Bradshaw *et al.*, 1994). Additionally, specific salivary proteins, such as MUC5B have been shown to only be utilisable as a nutrient source for bacteria when degraded by a complex of oral bacteria, rather than individual isolates from the same complex (Wickström and Svensäter, 2008).

The microbiome found within human saliva is distinct from the microbiomes of other oral structures, such as the tongue, tonsils, throat, and gingiva. Because of its ease of sampling, saliva has been one of the most widely studied oral features in humans. The microbiome of saliva is dominated by the Firmicutes, Bacteroidetes, Proteobacteria, Fusobacteria, and Actinobacteria phyla, with the *Streptococcus*, *Veillonella*, *Prevotella*, *Neisseria*, and *Fusobacterium* genera accounting for the majority of the microbiome found using culture-independent sequencing (Segata *et al.*, 2012). Interestingly, analysis of the salivary microbiome has shown there to be a high degree of individual differences within populations, with no observable geographical effect on its structure and composition. Additionally, variations within the salivary microbiome of a population appear to closely align with the degree of genetic differences that are commonly found within a population, approximately 13.5% (Nasidze *et al.*, 2009). Following this trend, longitudinal samples of the salivary microbiome of monozygotic and dizygotic twins suggest that the environment is more important than host genetics whilst shaping its structure and composition (Stahringer *et al.*, 2012).

4.1.3 | The Oral Metabolome

Saliva is around 99.5% water, containing a number of important biomolecules that aide in digestion and other bodily functions. Many of these biomolecules have origins in the circulatory system and for this reason, and because it is an easily accessible biofluid, with a non-invasive collection method, it is considered a potentially useful source of metabolomic biomarkers for human disease (Matias *et al.*, 2012). Primarily, saliva has been suggested as a novel biomarker pool for oral, head, and neck cancers, due to its localisation at the site of carcinogenesis. Indeed, metabolomics has identified altered levels of valine, lactic acid, and phenylalanine as being indicative of oral cancer, with sensitivity and specificity

frequencies above 80% (Wei *et al.*, 2011). Additionally, saliva has been shown to have promise for the detection of cancers not localised to the oral cavity, but nevertheless detectable because many salivary metabolites are derived from the circulatory system. Cancers that have shown particular promise through the identification of metabolome-derived biomarkers include those of the breast, pancreas, and lungs (Sugimoto *et al.*, 2010; Bonne and Wong, 2012).

Saliva is a promising biofluid for the identification of metabolome-derived markers for a number of diseases, including obesity (Matias *et al.*, 2012), dental caries (Yang *et al.*, 2012) and oral, breast, and pancreatic cancers (Sugimoto *et al.*, 2010). However, the oral metabolome is still relatively poorly understood in terms of gender and lifestyle-associated differences. Additionally, differences between stimulated and non-stimulated saliva could have important implications in the applicability of metabolome-derived disease biomarkers to clinical use (Neyraud *et al.*, 2012). Although limited to the use of ^{1}H NMR spectroscopy, metabolome differences between stimulated and non-stimulated saliva have been shown, particularly that metabolite concentrations are higher in non-stimulated saliva (Takeda *et al.*, 2009). This is likely a dilution effect of stimulated saliva containing a higher percentage of water than non-stimulated. Levels of fatty acids have also been shown to be increased in stimulated over non-stimulated saliva, reinforcing the importance of potential differences in stimulated and non-stimulated saliva when evaluating candidate disease biomarkers (Neyraud *et al.*, 2012). Furthermore, gender differences have been shown in the salivary metabolome, with a number of metabolites significantly higher in concentration in male saliva compared to female saliva, including acetate, glycine, lactate, pyruvate, and taurine. Changes in the salivary metabolome associated with tobacco smoking have also been shown, with metabolites including citrate, pyruvate, and sucrose showing significantly higher levels in male smokers compared to male non-smokers (Takeda *et al.*, 2009). As discussed previously, NMR is a useful technology for metabolome profiling, but it is limited in its detection sensitivity of low concentration metabolites. To date, limited work has been completed using mass spectrometry approaches to understand the salivary metabolome, and how it changes in association with physiological, environmental, and temporal changes.

Although saliva has received a large amount of attention in regards to its metabolome, and changes associated with a number of factors including disease, it is by no means the only feature of the oral cavity analysed through metabolomic techniques. For example, the metabolome of supragingival plaques, the site that is affected by dental caries, has been investigated through capillary electrophoresis mass spectrometry. This showed the close interplay between the metabolomes of the supragingival plaque and inhabiting oral bacteria, including the genera *Streptococcus* and *Actinomyces*, after the *in vitro* addition of glucose (Takahashi, Washio and Mayanagi, 2010). Additional work on the effects of glucose addition to microbial biofilms in the oral cavity, has further helped to elucidate the mechanistic role of the oral microbiome in disease (Washio, Mayanagi and Takahashi, 2010).

4.1.4 | Seasonal and Temporal Changes in the Human Body

The regulation of the human body in response to, or in anticipation of, changing environmental conditions is an evolutionary advantage; allowing for physiological and behavioural changes to occur. Physiological and behavioural responses, including weight and reproductive changes, in relation to seasonal alterations are well established in mammals, with melatonin the responsible hormone (Barrett and Bolborea, 2012). Melatonin has also been shown to be responsible for seasonal changes in the human immune system, namely cytokine production, neutrophil activity, and the differentiation and proliferation of lymphocytes (Klink *et al.*, 2012).

Hormonal-driven seasonal changes are not solely confined to the actions of melatonin, with a large number of human hormones displaying altered seasonal levels, which could have important implications on the human immune system. It is well established that vitamin D_3 is linked to immune function, and variations in its production can be largely contributed to ultraviolet light exposure. Recently, changes in vitamin D_3 levels in line with seasonal changes, are also associated with altered human peripheral T cell compartment function, and could explain some of the seasonal variation seen in human immune function (Khoo *et al.*, 2012). Additionally, testosterone levels have been shown, in both men and women, to vary considerably over a one year period, with a peak level being reached in the autumn

months, which can be twice the level of testosterone reached at the lowest level in the summer months (Stanton, Mullette-Gillman and Huettel, 2011). Temporal variations are not only observed in human hormonal levels, and fluctuations can be over short, medium, and long term periods. The human metabolome, for example, has also been shown to have a circadian rhythm, over a 24 hour period. Mass spectrometry of blood plasma and saliva samples, taken over a 40 hour period, suggested that approximately 15% of human metabolites are under circadian control (Dallmann *et al.*, 2012).

The human microbiome is closely linked to the physiological state of the host, and the state of the immune system in particular can have substantial effects on its structure and function. Early work on temporal and spatial differences in the human microbiome, from several body sites, found that spatial differences were significantly more substantial than temporal differences, though samples were only collected over a small time period, with the first and last collection separated by four months (Costello *et al.*, 2009). The human gut microbiome has recently been shown to drive hormone-dependent regulation of autoimmunity in mice, with male gut microbiome transplantations to females able to increase testosterone levels and affect progression of type 1 diabetes (Markle *et al.*, 2013). The microbiome of human milk has also been shown to have a temporal element to its composition, which may also be altered by hormonal levels present during birth (Cabrera-Rubio *et al.*, 2012a).

To date, the human microbiome and metabolome have tended to be studied separately, and at single time points, or at time points over multiple years. Seasonal changes in the human salivary microbiome and metabolome have not been measured. With increasing interest in the use of saliva-derived microbiome and metabolome biomarkers in a range of diseases, from oral conditions such as dental caries to cancers of peripheral organs, it is important to establish the degree of seasonal variation present to aide in the validation of such approaches. Additionally, it is well established that rates of upper respiratory tract infections is highest in the winter months, which may be linked to reduce immune function (Jones *et al.*, 2014). Consequently, seasonal variability in immune function may also lead to seasonal variability in the salivary microbiome.

4.1.5 | Aims and Objectives of Chapter

Establishing the temporal variability of the human microbiome and metabolome could be an important step in validating our understanding in terms of their relationship with health and disease. To date, this has not been accomplished over an annual period with frequent and regular sampling periods. In this portion of work, human saliva, which is arguably one of the easiest human biofluids to sample, was collected from 40 participants over a 12 month period, with collections every two months. Using these samples, the following will be determined to establish the temporal variability of the human salivary microbiome and metabolome:

1) Changes in total bacterial load over the 12 month period, using quantitative PCR.

2) Using a subgroup of ten participants, changes in taxonomic composition of the salivary microbiome over a 12 month period through 16S rRNA amplicon sequencing.

3) Metabolomic fingerprinting of saliva to monitor whether any observed bacterial changes are associated with changes in the biological composition of the saliva.

4.2 | Materials and Methods

This study received ethical approval from the Aberystwyth University Research Ethics Committee. Informed consent was obtained from all participants at least 24 hours before the first sample was taken, and additional consent forms were obtained before each subsequent sample was donated. All participant information obtained was link anonymised prior to subsequent data analysis.

4.2.1 | Participant Recruitment and Sampling

Stimulated saliva samples were obtained from 40 participants, consisting of staff and students at Aberystwyth University, over a one year period, from October 2012 to October 2013. During this period, a total of seven samples were collected every two months, each over a twelve day period, including October 2012 (10/09/2012 to 21/09/2012), December 2012 (10/12/2012 to 21/12/2012), February 2013 (11/02/2013 to 22/02/2013), April 2013 (08/04/2013 to 19/04/2013), June 2013 (10/06/2014 to 21/06/2014), August 2013 (12/08/2013 to 23/08/2013), and October 2013 (14/10/2013 to 25/10/2013). Participants donated 5 mL of saliva into a sterile 50 mL centrifuge tube, and stimulated saliva through jaw movement if necessary. Participants were not restricted in eating or drinking prior to donating a saliva sample. At the same time, information on oral hygiene practice, antibiotic use, smoking history and diet was collected.

4.2.2 | Sample Processing and Total Genomic DNA Extraction

All saliva samples were normalised to ensure 5 mL volume and underwent centrifugation at 10,000 x g for 20 minutes, at 4°C, after which 2 mL of the saliva supernatant was transferred to a PCR grade microcentrifuge tube. The remaining saliva supernatant was removed and destroyed, and the saliva pellet transferred to a PCR grade microcentrifuge tube. The saliva pellet was stored at -80°C for a maximum of seven days, after which it underwent DNA extraction. The saliva supernatant was stored at -80°C until collection at all sampling time points had been completed. Subsequent metabolomic fingerprinting was completed at one time point.

Total genomic DNA was extracted from 200 µL of the saliva pellet using a FastDNA SPIN kit for soil (MP Biomedical, Santa Ana, USA) following manufacturer's instructions. Bead beating was carried out in a FastPrep-24 machine (MP Biomedical) with three cycles at speed setting 6.0 for seconds, with cooling on ice for 60 seconds between cycles. Genomic DNA was eluted with 50 µL of DES and dsDNA concentration determined, in duplicate, using 2 µL on the Epoch spectrometer system (BioTek).

4.2.3 | 16S rRNA Quantitative PCR

Quantitative PCR was carried out on neat extracted DNA against standards created by amplifying the 16S rRNA gene of five randomly selected October 2012 samples. This was completed through amplification of the 16S rRNA gene in a 20 µL reaction volume consisting of 10 µL of 2 x BioMix (BioLine), 0.25 µL each of 27f (5'-AGA GTT TGA TCC TGG CTC AG-3') and 1389r (5'-ACG GGC GGT GTG TAC AAG-3') primers (Hongoh, Ohkuma and Kudo, 2003) to give a final concentration of 500 nM, 1 µL of neat extracted DNA, and 9.5 µL of PCR Grade Water (Roche). The reaction volumes were then subjected to PCR consisting of 94°C for two minutes, 30 cycles of 94°C for 45 seconds, 55°C for 45 seconds, and 72°C for 90 seconds, followed by a final elongation step of 72°C for seven minutes. The resulting PCR products were combined and purified using an Isolate II PCR and Gel Extraction purification kit (BioLine), following manufacturer's instructions, and quantified with an Epoch spectrometer as previously described. The resulting DNA concentration was used to estimate the total number of 16S rRNA gene copies and serial dilutions of 10^{10}, 10^{8}, 10^{6}, 10^{4}, 10^{2}, and 10^{0} made.

Quantitative PCR was completed on neat extracted DNA against standards with each reaction completed in 25 µL volumes, each consisting of 12.5 µL 2 x SYBR Green Mastermix (Life Technologies), 0.25 µL of each EubF1 (5'-GTG STG CAY GGY TGT CGT CA-3') and EubR1 (5'-ACG TCR TCC MCA CCT TCC TC-3') primer (Maeda *et al.*, 2003), in a final concentration of 400 nM, 11 µL of PCR Grade Water (Roche) and 1 µL of neat DNA extract. Reactions were run using a C100 thermal cycler (BioRad, Hercules, USA) and CFX96 optical detector (BioRad), with data captured using CFX Manager software (BioRad), under conditions of 95°C for ten minutes, 40 cycles of 95°C for 15 seconds and 60°C for 60 seconds, followed

by a melt curve consisting of a temperature gradient of 60°C to 95°C in 0.5°C increments, each for five seconds. Where shown, *P* values represent the significance of one-way ANOVA tests.

4.2.4 | Selection of Participants for 16S rRNA Amplicon Sequencing

Of the 40 recruited participants in this study, a subgroup of ten was selected for 16S rRNA amplicon sequencing of all seven monthly samples collected. This subgroup was selected based on supporting information given at each bi-monthly sample, with a view to selecting a group of participants with minimal differences. Participants were selected for 16S rRNA amplicon sequencing based on oral hygiene practices (no history of mouthwash but a history of flossing at least weekly), smoking history (no current smokers and past smokers with a cessation period greater than ten years), allergen history (no asthma or hay fever), diet (meat-eaters only), antibiotic exposure (no antibiotic use within sampling period and six months prior to start), with no restriction on age or gender.

4.2.5 | 16S rRNA Amplicon Preparation

Sequencing of the 16S rRNA gene was carried out via amplification of the V3 to V4 region and subsequent amplicon sequencing on the Illumina MiSeq platform. Firstly, the V3 to V4 region of the 16S rRNA gene was amplified through duplicate PCR with locus specific primers, alongside negative water controls. In a 25 µL reaction volume, 12.5 ng of extracted DNA, or 2.5 µL of PCR grade water for negative controls, was added to 12.5 µL of 2 x Accuzyme Mix (BioLine) and 5 µL each of a 1 µM concentration of 319f primer (5'– CCT ACG GGN GGC WGC AG–3') with Illumina forward overhang adapter sequence (5' – TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG-3') and 806r primer (5'–GAC TAC HVG GGT ATC TAA TCC–3') with Illumina reverse overhang adapter sequence (5'- GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G-3') as detailed by Klindworth *et al.*, (2013). The reaction mix underwent PCR consisting of 95°C for three minutes, followed by 25 cycles each of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by a final elongation step of 72°C for five minutes. Each duplicate PCR volume was confirmed through visualisation on a 2% agarose gel, after being run for 120 minutes at 100 volts (≤

80 mA) in 1% tris base, acetic acid, and ethylenediaminetetraacetic acid (TAE) buffer. Following confirmation of PCR success, corresponding reaction volumes were combined and purified using an Isolate II PCR and Gel Extraction kit (BioLine), following manufacturer's instructions, with elution into 20 µL of kit buffer. Following purification, a second PCR was completed to attach Illumina adaptors to amplified products to allow for multiplexed amplicon sequencing on the Illumina MiSeq platform. To make a final reaction volume of 25 µL, 2.5 µL of purified PCR product from the previous step was added to 12.5 µL of 2 x Accuzyme Mix (BioLine), 5 µL of PCR Grade Water (Roche) and 2.5 µL each of the relevant Nextera XT Index Primer 1 (N7##) and Nextera XT Index Primer 2 (S5##) (Illumina) as detailed in Chapter 4 Appendix, Supplementary Table 4.1. The reaction mix underwent a limited cycle PCR consisting of 95°C for three minutes, eight cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by a final elongation step of 72°C for five minutes. To remove non-combined adaptors, the entire reaction volume was run on a 2% agarose gel, in TAE buffer, for 120 minutes at 100 volts (≤ 80 mA). The gel was visualised using a DR195M Transilluminator (Clare Chemical Research, Colorado, USA) and each PCR product excised using a sterile scalpel blade. PCR products were purified using an Isolate II PCR and Gel Extraction kit (BioLine), following manufacturer's instructions, with elution into 20 µL of kit buffer. Purified PCR products were quantified using a Quant-iT dsDNA High Sensitivity assay kit and a Qubit fluorometer (Life Technologies).

4.2.6 | 16S rRNA Amplicon Sequencing and Analysis

Individual sample libraries were pooled together in equimolar concentration and sequenced, along with 20% PhiX DNA as a control for low diversity, on the Illumina MiSeq platform using MiSeq v3 reagents for a 2 x 300 bp run at the IBERS Translational Genomics Facility, Aberystwyth University. After sequencing, sample reads were demultiplexed and trimmed for quality, with overlapping reads merged using FLASH (Magoč and Salzberg, 2011). Merged reads were analysed using the MG-RAST metagenomics analysis pipeline (Meyer *et al.*, 2008). Taxonomic alignments of sequences was completed using the Ribosomal Database Project (Cole *et al.*, 2009) facility, with only those sequences with a minimum alignment identity of 97%, maximum e-value of 1×10^{-5}, and a minimum alignment cut-off of 15 being used.

Sequences were analysed using the PCA facility and α-diversity measure, based on the Shannon diversity index, within MG-RAST, and exported into Microsoft Excel 2010 and MINITAB 14 for further analysis, with multivariate statistics completed using MetaboAnalyst 2.0 (Xia *et al.*, 2012). Sequence numbers for each sample were normalised as a percentage composition of the total volume of sequences for each taxonomic level of classification. Where shown, *P* values represent the significance of one-way ANOVA tests. All sequence files are available under the MG-RAST project ID 11549: 'Charting Temporal Variability in the Salivary Microbiome'. Raw sequence files were deposited at the European Nucleotide Archive under primary accession number PRJEB9010 and secondary accession number ERP010064.

4.2.7 | LTQ-MS Metabolomic Fingerprinting

Saliva supernatant was thawed at 4°C and 200 µL transferred to a sterile 2 mL microcentrifuge tube, to which 30 mg of ≤ 106 µM acetone-washed glass beads (Sigma-Aldrich, Dorset, UK) were added. To this, 1520 µL of a solvent mix of HPLC grade methanol and chloroform, in a ratio of 4:1, was added. To homogenise the mixture, samples were placed on a vortex mixer for five seconds and then milled for 30 seconds at 30 Hz. Following this, samples were shaken for 20 minutes at 4°C and then stored at - 20°C for 20 minutes to precipitate protein. Samples then underwent centrifugation at 11 000 x g for six minutes at 4°C. The resulting supernatant was removed and transferred to a sterile 2 mL microcentrifuge tube. From this, 70 µL was transferred to an LTQ-MS vial and sealed. Samples were stored at -20 °C until run, in a randomised order using an autosampler, with tray temperature kept constant at 15°C. For each sample, 20 µL was injected into a flow volume of 60 µL per minute water-methanol, in a ratio of 70% water and 30% methanol, using a Surveyor liquid chromatography system (Thermo Scientific, MA, USA). Data acquisition for each individual sample was conducting, in alternating positive and negative ionisation mode, over four scan ranges (15-110 m/z, 100-220 m/z, 210-510 m/z, 500-1200 m/z) on an LTQ linear ion trap (Thermo Electron Corporation, CA, USA), with an acquisition time of five minutes. Individual metabolite m/z values were normalised as a percentage of the total ion count for each sample. Normalised abundances were subsequently analysed using MetaboAnalyst 2.0 (Xia *et al.*, 2012) and PyChem (Jarvis *et al.*, 2006).

4.2.8 | pH Measurements of Saliva

Measurements of the pH of saliva supernatant was carried out after processing of each saliva sample and using a B-212 Twin pH Meter (Horiba, Kyoto, Japan) after two point calibration using pH 7 and pH 4 buffers. For pH measurements, 200 µL of saliva supernatant, after thawing at 4°C, was used. After stabilisation of reading, pH value was recorded and the sensor washed with ultrapure water, and blotted dry. Data analysis was completed using Microsoft Excel 2010 and Minitab 14. Where shown, *P* values represent the significance of one-way ANOVA tests.

4.3 | Results

Saliva samples were collected from a total of 40 participants over a one year period, with sampling occurring over a two week period every two months, from October 2012 to October 2013. Participant information for the complete sample group is detailed in Table 4.1, alongside the characteristics of the sub-group of ten participants selected for 16S rRNA amplicon sequencing. The sequencing sub-group was specifically selected based on their lifestyle similarities. Full participant information is detailed in Chapter 4 Appendix, Supplementary Table 4.2.

TABLE 4.1 | Lifestyle History of Whole Sample Group and Sequencing Sub-Group
Group summaries of whole sample group (n=40) and sequencing sub-group (n=10). Relevant group means are shown alongside standard deviations, italicised in brackets. All information was self-reported.

		Whole Group	Sequencing Sub-Group
Age		41.75 (*13.14*)	44.90 (*14.86*)
Gender Ratio (Male : Female)		24 16	7 3
Smoking History	**Current**	4/40	0/10
	Smoking Pack Years	2.19 (*2.10*)	- -
	Past	9/40	2/10
	Smoking Pack Years	9.47 (*8.24*)	10.50 (*6.36*)
	Average Cessation (Years)	14.80 (*10.42*)	27.50 (*3.54*)
	Never	27/40	8/10
Asthma History		3/40	0/10
Hay Fever History		5/40	0/10
Oral Hygiene Practice	**Mouthwash Use**	19/40	0/10
	Antibacterial Mouthwash Use	17/19	0/10
	Manual Toothbrush Use	23/40	3/10
	Electric Toothbrush Use	17/40	7/10
	Flossing	26/40	10/10
	Flossing Frequency (Days)	3.46 (*2.39*)	2.90 (*2.64*)
Diet	*Meat Eater (1 – 3 days)*	9/40	4/10
	Meat Eater (4 – 7 days)	26/40	6/10
	Vegetarian	5/40	0/10

4.3.1 | Temporal Changes in Bacterial Load

The estimated bacterial load of the saliva was measured through qPCR, targeting the 16S rRNA gene, using population standards taken from five randomly selected October 2012 samples. Average estimated bacterial loads for all 40 participants, at each time point, are given in Figure 4.3a, alongside average individual changes from one time point to the next, and from October 2012 to October 2013, Figure 4.3b. One-way ANOVAs show that the February 2013 time point had a significantly ($P < 0.001$) higher estimated bacterial load than all other time points. Additionally, when individual changes in estimated bacterial load are calculated, changes can be seen with February 2013 to April 2013 and June 2013 to August 2013 changes showing a significant ($P < 0.001$) net decrease in an individual's estimated bacteria load. Overall, there was no net change in estimated bacterial load over the entire sampling period. However, substantial and significant, flux was observed over the time course; particularly between individual time points.

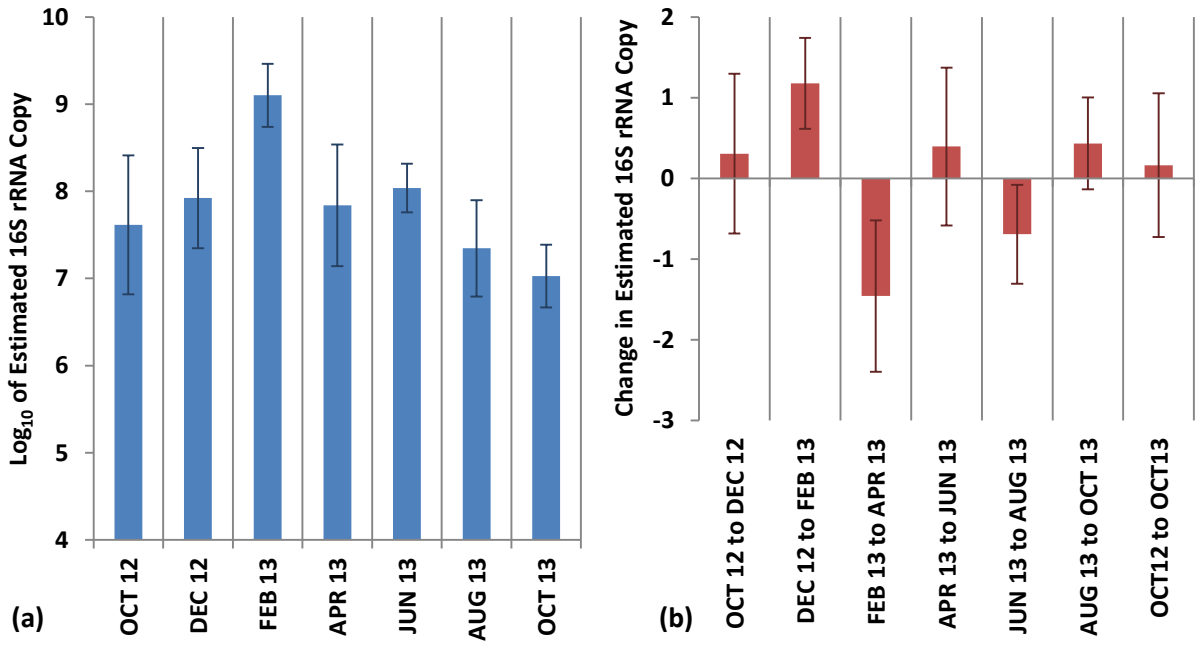


FIGURE 4.3 | Estimation of Salivary Bacterial Load
Estimated bacteria load was measured through qPCR targeting the 16S rRNA gene. Average estimated bacteria loads by (a) time point show a significantly ($P < 0.001$) higher bacterial load in February 2013 than at all other time points. Additionally, average (b) individual changes from one time point to the next show a significant ($P < 0.001$) level of flux, with net decreases shown only in the February 2013 to April and June 2013 to August 2013 time point. Error bars in figures show one standard deviation around the mean.

4.3.2 | Temporal Changes in Taxonomy of Salivary Microbiome

Amplicon sequencing statistics are detailed in Chapter 4 Appendix, Supplementary Table 4.3, and show no significant differences in total sequence base pairs by participant (P = 0.268), or month (P = 0.537), or total sequence number by participant (P = 0.247) or month (P = 0.542). However, sequence length by participant were significantly different (P < 0.001), with a range of approximately 15 bp. However, no such differences were seen in sequence length by month (P = 0.101). The GC content of sequences was also significantly different by participants (P < 0.001), but not by month (P = 0.896).

Using the MG-RAST online platform, principal component analysis was completed using taxonomy assigned by the RDP database, with an alignment similarity of 97%, and with other parameters at their

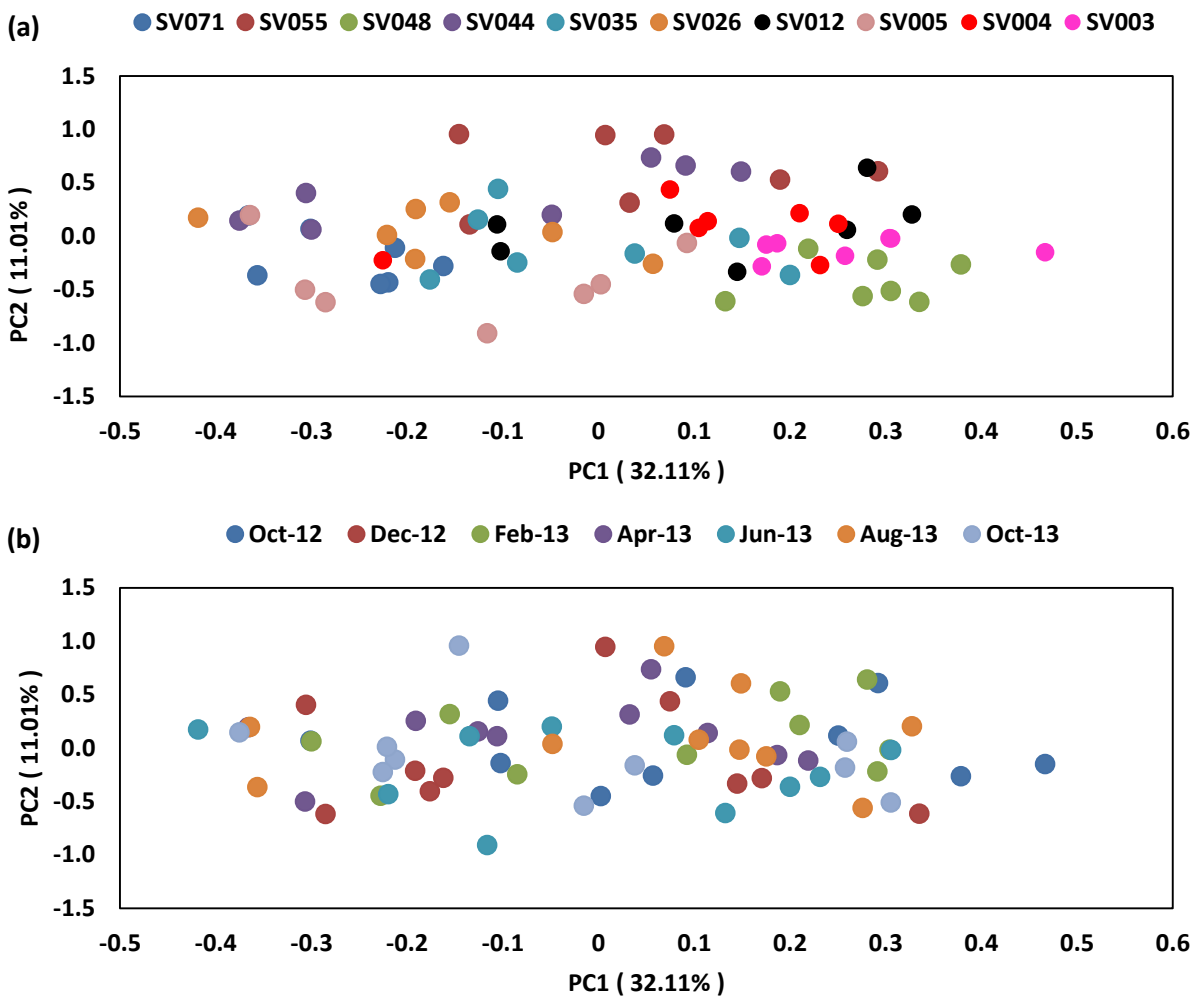


FIGURE 4.4 | Principal Component Analysis of 16S rRNA Taxonomy
Using the MG-RAST online platform, principal component analysis was completed, using taxonomy classifications from the RDP database, at a 97% similarity alignment, with other parameters at default. Resulting plots show partial separation by (a) participant, but not by (b) sampling month.

default values. Partial separation was seen by participants, Figure 4.4a, but not by month, Figure 4.4b. Separation by participant was notably stronger in certain participants, such as SV048, SV004, SV003 and SV055, compared to others.

Analysis of species diversity within a sample at each time point was calculated using the MG-RAST online platform. Averages of α-diversity are given in Figure 4.5 by (a) participant and (b) month. Significant differences were seen between participants ($P < 0.001$) but not between sampling months ($P = 0.801$).

From principal component analysis and α-diversity values, it is evident that the variation between participants is substantially, and significantly, greater than that seen between sampling time points, suggesting the presence of relative temporal stability within the salivary microbiome, in terms of taxonomic diversity. Although macro-level differences are not seen within the taxonomic diversity of the salivary microbiome, micro-level changes, at the genus level may nonetheless be evident. Using the MetaboAnalyst 2.0 online platform, one-way ANOVAs were completed to identify genera that may be significantly altered in their abundance over the sampling time course. The genera *Rhodococcus* ($P = 0.006$) and *Variovorax* ($P < 0.050$) were shown to have significantly different abundances over the time

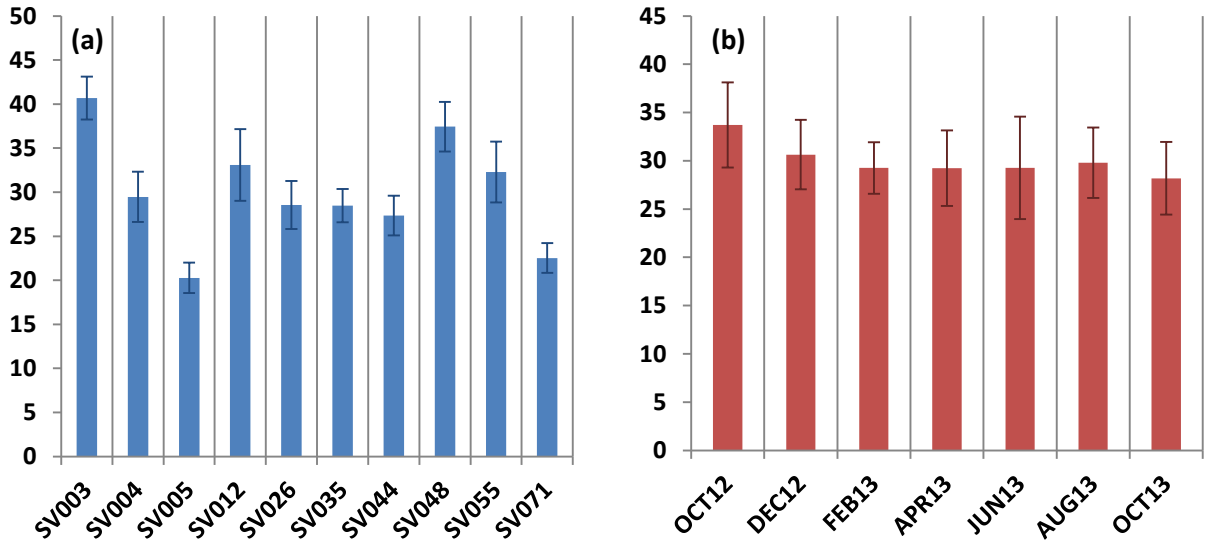


FIGURE 4.5 | α-Diversity Values by Participant and Month
Species diversity within a sample at each time point was calculated using the MG-RAST online platform, with averages of α-diversity given by (a) participants, and by (b) month. Significant ($P < 0.001$) differences were observed between participants, but not between sampling months ($P = 0.801$). Errors bars display one standard deviation around the mean.

course of sampling. However, both of these genera displayed very low abundance levels, and were present in less than 50% of all samples, with *Variovorax* only present in two samples. Therefore, it is likely that these significant values are statistical artefacts of the genus's low abundances, as they were not reproducible in additional statistical packages.

After establishing that individual differences in the taxonomic composition of the salivary microbiome are more significant than any seasonal effect which may exist, the phylum level differences between participants were established. The Firmicutes phylum was the largest of the phyla, although the number of unclassified sequences, with a suspected bacterial origin, contributed a substantial proportion of the total bacterial reads, up to 50% of reads in some samples. Within the phylum level of classification, Figure 4.6, the Actinobacteria ($P < 0.001$), Bacteroidetes ($P < 0.001$), Firmicutes ($P = 0.008$), Fusobacteria

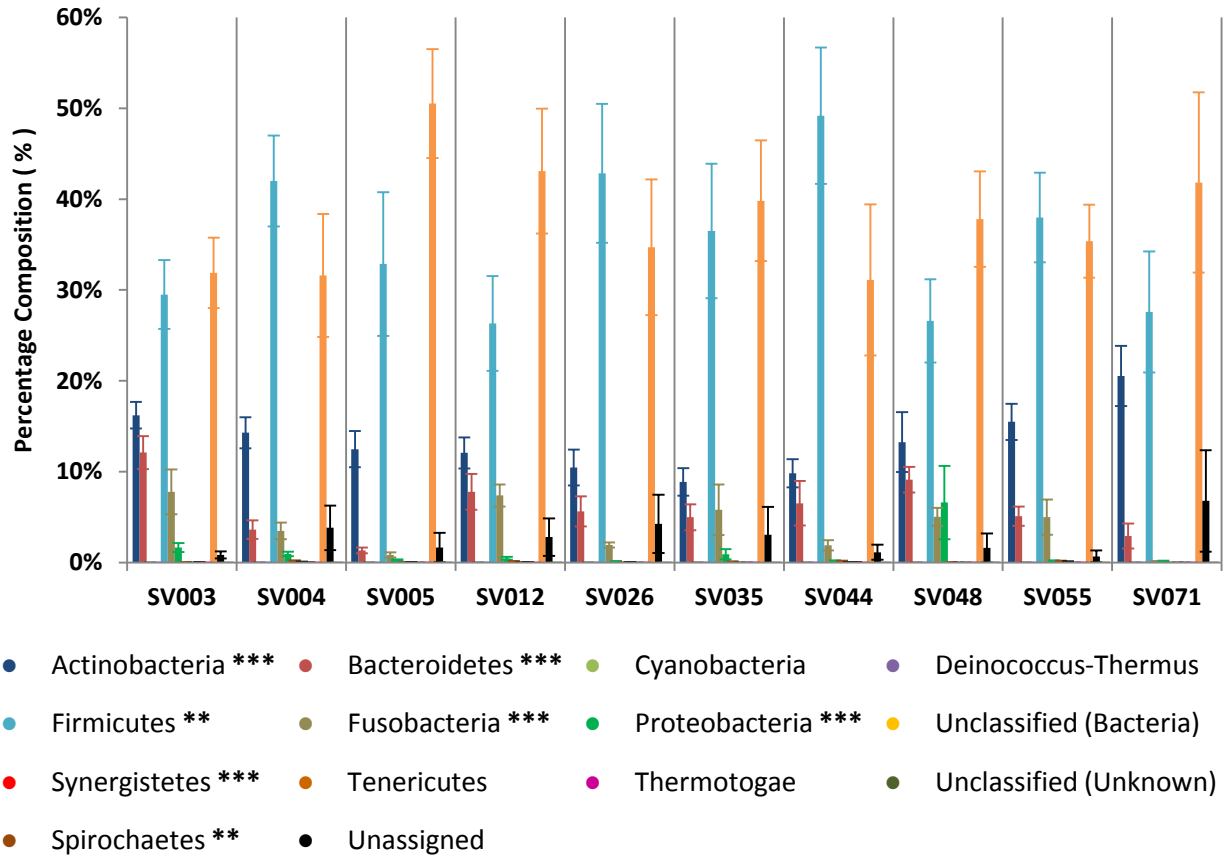


FIGURE 4.6 | Average Phylum Level Taxonomy for 16S rRNA Sequencing Sub-Group
Individual differences have been shown to be more substantial in determining the taxonomic composition of the salivary microbiome than any temporal or seasonal factors. At the phylum level of classification, these individual differences are pronounced, with a number of phyla displaying significantly different abundances between participants. Significance thresholds, as determined through one-way ANOVAs, are indicated in figure legend (*** = $P < 0.001$; ** = $P < 0.01$).

($P < 0.001$), Proteobacteria ($P < 0.001$), Synergistetes ($P < 0.001$), and Spirochaetes ($P = 0.003$) were shown to be significantly different between participants.

4.3.3 | Temporal Changes in Salivary pH

The pH of any environment can be an important factor in the ability of microorganisms to inhabit and grow. As with estimated bacteria load, the pH of saliva samples was measured at each time point, and the time point average, Figure 4.7a, and average individual time point difference, Figure 4.7b, was calculated. Salivary pH was shown to be significantly ($P = 0.003$) higher in December 2012 compared to October 2012 and February 2013 time points. As with estimated bacterial load, an individual's salivary pH levels were also shown to be in flux throughout the sampling period. Although over the one year period there was no net overall change, there were significant ($P < 0.001$) changes from one point to the next, Figure 4.7b.

In relation to other measured variables, pH was shown to have no significant ($P = 0.219$) relationship

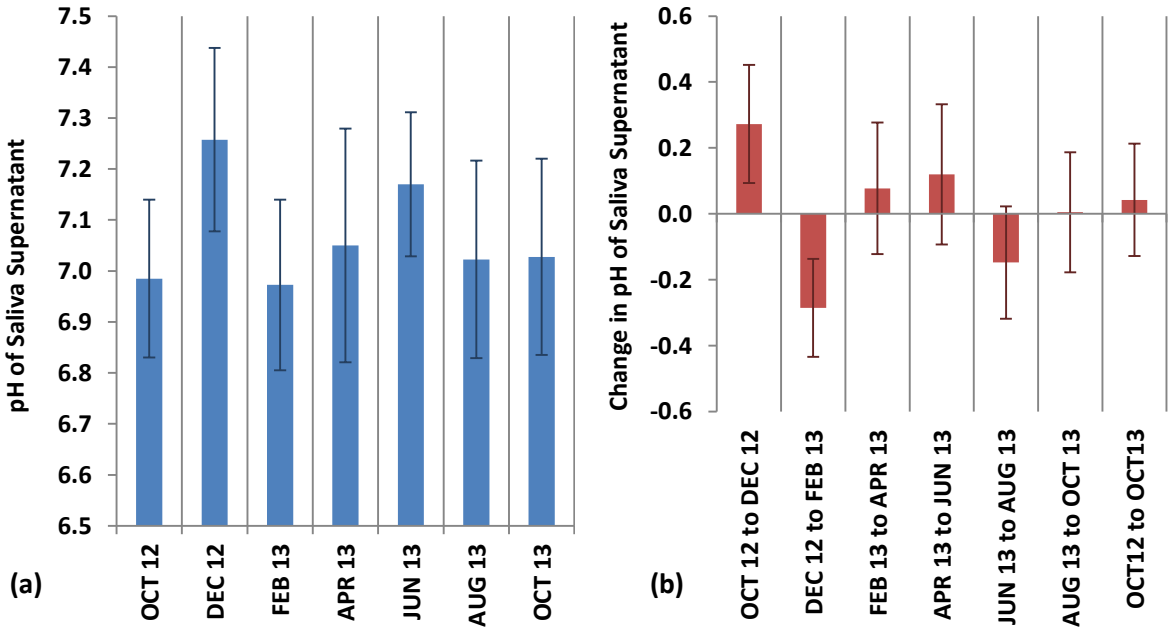


FIGURE 4.7 | Salivary pH Levels
Salivary pH average for each (a) time point, and (b) individual changes between each time point, were measured. The December 2012 time point was shown to have a significantly ($P = 0.003$) higher pH than the October 2012 and February 2013 time points only. Individual differences between time points were significant ($P < 0.001$), though there was no overall net change over the entire sampling period. Error bars shown are one standard deviation around the mean.

with estimated bacterial load, or any measured LTQ-negative metabolite. However, salivary pH levels were shown to have a small but significant positive correlation with α-diversity values (R^2 = 7.8%, *P* = 0.019).

4.3.4 | Temporal Changes in Salivary Metabolome

The chemical composition of saliva may be an important component in determining the taxonomic composition of the salivary microbiome. For all 40 participants, at all time points, metabolomic profiles were generated using negative mode LTQ-MS. Principal component analysis, Figure 4.8, showed no separation of profiles based on sampling time point. Additional modelling, not shown, also showed no separation between participants based on negative mode LTQ-MS metabolomic profiles.

Additional analysis of all negative mode LTQ-MS metabolites showed no significant correlations between estimated bacterial load nor salivary pH levels. However, a number of metabolites did exhibit significant correlations with seven bacterial genera, Table 4.2, using the 70 samples that underwent 16S rRNA amplicon sequencing. A total of 11 correlations were observed with significant regression values above a R^2 value of 90%. However, all significant correlations were observed between negative mode

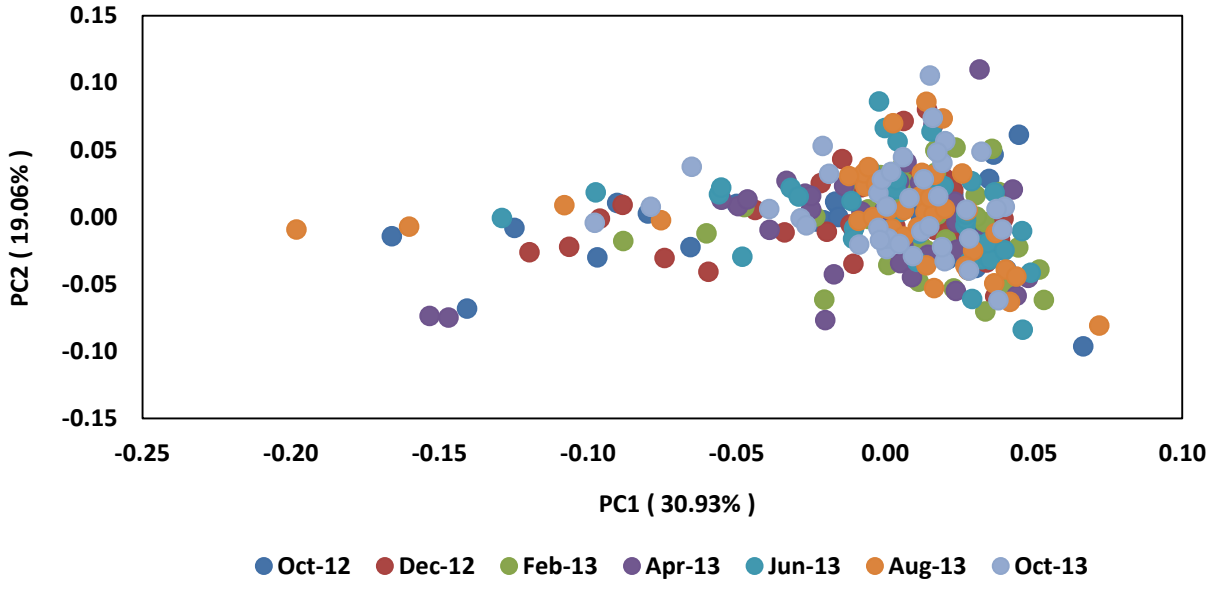


FIGURE 4.8 | Principal Component Analysis of Negative Mode LTQ-MS Metabolites
Principal component analysis of negative mode LTQ-MS metabolite profiles was completed in PyChem, with samples classified by collection time point. Minimal variation was observed between samples, and no separation was observed between time points.

TABLE 4.2 | Regression Analysis of Genera Taxonomy and Negative Mode LTQ-MS Metabolites
Significant correlations were observed between seven bacterial genera and a total of seven metabolites derived from negative mode LTQ-MS fingerprinting. Genus count refers to the number of samples that underwent 16S rRNA amplicon sequencing in which sequences aligned to that genus are found.

Genus	Metabolite (m/z)	R^2 (%)	R^2-adjusted (%)	*P* Value	Genus Count
Janibacter	269	92.4	92.3	< 0.001	1 / 70
	431	97.3	97.2	< 0.001	1 / 70
	432	96.4	96.4	< 0.001	1 / 70
	539	96.5	96.5	< 0.001	1 / 70
	593	91.2	91.1	< 0.001	1 / 70
Cellulophaga	39.18	99.8	99.8	< 0.001	1 / 70
Bradyrhizobium	161.9	98.7	98.7	< 0.001	1 / 70
Desulfovirga	161.9	98.7	98.7	< 0.001	1 / 70
Alkalispirillum	161.9	98.7	98.7	< 0.001	1 / 70
Klebsiella	161.9	98.7	98.7	< 0.001	1 / 70
Acholeplasma	39.18	99.8	99.8	< 0.001	1 / 70

LTQ-MS metabolites and genera which were found in only one of the 70 samples that underwent 16S rRNA amplicon sequencing.

4.4 | Discussion

The human microbiome and metabolome are important components of homeostasis, and their dysbiosis has been linked to disease in both cause and effect. In understanding their role, the temporal variability of the human microbiome and metabolome has not been definitively established. Due to sampling ease, saliva has been suggested as a source of biomarkers for a number of diseases, including oral, breast and pancreatic cancers (Sugimoto *et al.*, 2010), obesity (Matias *et al.*, 2012), and dental caries (Yang *et al.*, 2012). Whether the human salivary microbiome and metabolome displays temporal variability is likely to determine the usefulness of disease biomarkers derived from the microbiome or metabolome. In this portion of work, the salivary microbiome and metabolome of 40 people was sampled over a one year period, every two months. For all participants, the estimated salivary microbial load and pH was measured, and metabolomic profiles constructed through negative mode LTQ-MS. Additionally, a group of ten participants, selected because of their lifestyle similarities, underwent 16S rRNA amplicon sequencing to determine the taxonomic composition of their salivary microbiome.

4.4.1 | Temporal Stability of the Salivary Microbiome

The bacterial load of human saliva has been suggested as an *in vivo* marker of immunity, and previous work has shown an increase in salivary bacterial load over the winter months (Jones *et al.*, 2014). However, the focus of this study was not the temporal stability of the microbiome, and looked only at the salivary microbiome of physically-active males. Nevertheless, a similar pattern of salivary bacterial load was seen in this portion of work. The sampling period February 2013 displayed the highest level of estimated salivary bacterial load. When the change in bacterial load is calculated for an individual, it can be seen that bacterial load appeared relatively stable over the 12 month sampling period. However, there are considerable net increases and decreases from one time point to the next. It has been suggested that there may be a link between salivary bacterial load and *de novo* plaque formation (Dahan *et al.*, 2004), although this has been disputed by others (Rowshani, Timmerman and Van der Velden, 2004). Salivary bacterial load has also been shown to not be associated with common dental

conditions such as gingivitis and periodontal disease (Mantilla Gomez *et al.*, 2001). However, these studies relied on the use of culture-dependent techniques such as counting of colony forming units. It may be that there is no link between the bacterial load of cultureable bacteria and common dental diseases, but a link with difficult-to-culture bacteria cannot be dismissed.

No relationships were seen between estimated salivary bacterial load and salivary pH level, α-diversity of the salivary microbiome, nor any negative mode LTQ-MS metabolites. This suggests that the variable, or variables, associated with changes in salivary bacterial load were not measured in this portion of work. Salivary bacterial load has been suggested as an *in vivo* measure of immunity and it may be that markers of the human immune system, such as immunoglobulin factors, may show an association. Although not measured in this portion of work, markers for the human immune system were measured by Jones *et al.*, (2014) who did not find any association with salivary bacterial load.

The focus of many microbiome studies is in characterising its taxonomic composition, and there are multiple methods which can be used to accomplish this. In this portion of work, the culture-independent sequencing of 16S rRNA amplicons was used, allowing for taxonomic classifications to the genus level. Analysis of the salivary microbiome through principal component analysis, Figure 4.4, showed that individual samples separated by participants rather than collection time point. The samples of particular participants appeared to cluster more closely than others, suggesting that individual differences between participants are the greatest determinant of the salivary microbiome. The measure of species diversity used in this portion of work, α-diversity, was also shown to be determined more by participant than by sampling time point, Figure 4.5. Significant differences were observed between participants, but not sampling time points. The standard deviations of α-diversity values for participants suggest that species diversity within the salivary microbiome is constant throughout the sampling period. Interestingly, no relationship between α-diversity and estimated salivary bacterial load was observed, suggesting that the increase in salivary bacterial load seen in the February 2013 time point is an equal increase in all bacteria, rather than specific taxa. A significant correlation was however observed

between α-diversity and salivary pH, though only 7.8% of variation in diversity was explained. Due to its impact on enzyme function, salivary pH is an important determinant in bacterial colonisation, and growth. The positive correlation between salivary pH and bacterial diversity suggests that as saliva become increasingly acidic, the range of bacteria able to tolerate these conditions decreases. Lower salivary pH levels have been linked to oral diseases, such as dental caries (Humphrey and Williamson, 2001), and it could be that a corresponding reduction in bacterial diversity is an additional factor.

In regards to differences between sampling time points within individual genera, only *Rhodococcus* and *Variovorax* were significantly different between time points using the MetaboAnalyst 2.0 online platform. However, these genera were present in only a minority of samples, with *Variovorax* present only in two. Furthermore, the significant differences shown in the MetaboAnalyst 2.0 platform were not reproducible in additional statistical packages. Therefore, it was determined that the significance displayed was an artefact of their low abundance. As with bacterial diversity, significant differences were only evident between participants, rather than sampling time point. At the phylum level of classification, seven phyla were seen to have significantly different abundances between participants. The large number of unclassified bacterial sequences evident in samples, with an average range of between 30% and 50%, is noteworthy. It may be possible that significant differences are indeed present within the taxonomic composition of the salivary microbiome, but that these differences exist within poorly defined taxa.

The taxonomic composition and diversity of the salivary microbiome in this portion of work appeared to be determined by individual differences, rather than temporal changes over the one year sampling period. This is in line with other work into the human salivary and oral cavity microbiome, although over different temporal periods. For example, Stahringer *et al.*, (2012) found that the human salivary microbiome appears remarkably stable once in adulthood, which may be as a result of a stabilisation in diet, oral hygiene, and other lifestyle factors. Over a shorter time period, namely three months, the oral cavity and other body sites displayed a high degree of temporal stability (Costello *et al.*, 2009). The

sampling time period in this portion of work appears to be unique within the published literature, and is the first to suggest that comparing salivary microbiome composition, but not bacterial load, from any time point within the year is valid.

It has been suggested that a core microbiome in healthy individuals exists in the form of metabolic function, rather than taxonomic composition (Turnbaugh *et al.*, 2009). Through sequencing of the 16S rRNA gene in this portion of work, only the taxonomic make-up of the salivary microbiome could be established. To establish the functional capacity of the salivary microbiome, metagenomic sequencing of the entire DNA found within a sample would be required. This method of sequencing however requires substantial resources which were not available to this project. Additionally, metagenomic sequencing allows for the assignment of species or even strain-level taxonomy, and it may be that temporal variation exists within these classifications (Weinstock, 2012).

One of the current limitations of microbiome research is the common focus of only characterising the bacterial component, as was the case in this portion of work. This is arguably associated with the ease of classifying bacteria based on sequencing of the 16S rRNA gene, and although the same technique can be used for the eukaryotic 18S rRNA gene, it is complicated by the high abundance of host sequences, particularly in humans. The lack of curated databases for eukaryotic microbes is also an issue, which complicates the biological interpretation of sequence data (Grice and Segre, 2012). As with the eukaryotic microbiome, the viral microbiome is also a neglected area of human microbiome research. In saliva, bacteriophages have been shown to be a significant reservoir of bacterial pathogenic gene function, which is temporally instable. However, it may be that the virome is constantly in a state of flux, rather than reflecting a composition associated with the time of year (Pride *et al.*, 2012).

4.4.2 | Temporal Stability of the Salivary Metabolome

Based on negative mode LTQ-MS profiles, no detectable temporal differences were evident within the salivary metabolome of all 40 participants in this portion of work, and additionally no differences were

evident between participants, based on principal component analysis. In addition to consumed food present within the oral cavity, saliva acts as the main growth medium for the human salivary microbiome. It is therefore expected that temporal stability in the human salivary microbiome is reflected in the salivary metabolome. However, the differences evident between participants in the salivary microbiome do not appear to be reflected in the salivary metabolome. This suggests that additional factors than those found in the salivary metabolome may be the causative factor in determining the taxonomic composition of a participant's salivary microbiome, such as immunoglobulin factors (Jones *et al.*, 2014) or salivary antimicrobial compounds (Peters, Shirtliff and Jabra-Rizk, 2010).

Saliva has been extensively studied using a variety of metabolomic techniques, with the focus being on identifying disease biomarkers. However, minimal work has evaluated the temporal stability of the salivary metabolome over a prolonged period of time. The collection time of saliva, namely morning and afternoon, has been shown to influence the salivary metabolome; though differences in age was believed to be a significant contributory factor to the observed changes (Sugimoto *et al.*, 2012). The human circadian rhythm has also been suggested to influence the salivary metabolome, though only approximately 15% of metabolites were under direct or indirect control by the circadian clock mechanism (Dallmann *et al.*, 2012). These findings suggest that the salivary metabolome is in a continuous state of flux, with its composition determined by more than one factor.

Through comparison of individual negative mode LTQ-MS metabolites and genus-level taxonomic assignments for the ten participants selected for the sub-group used in 16S rRNA amplicon sequencing, a number of significant correlations were observed. A total of seven bacterial genera displayed a significant relationship with at least one of seven negative mode LTQ-MS metabolites, which explained over 90% of the variation observed. However, these genera could be considered low abundance, as they appear in only one of the 70 samples that underwent 16S rRNA amplicon sequencing. It may be that these significant regression analyses are statistical anomalies of the low abundance, but it may also be that these metabolites are essential for the growth of these genera. The combination of microbiome

and metabolome profiling is not common, though several studies have attempted small scale proof-of-concept studies. Celiac disease, for example, has been studied through both of these methods, with metabolomics providing biomarkers for the disease, and microbiomics providing a possible mechanism for the disease (Sellitto *et al.*, 2012). The combination of microbiome and metabolome profiling in humans has the potential to unravel the mechanisms behind disease, and to provide possible targets for treatment. Nevertheless, these two techniques only account for a portion of the approaches needed for a systems biology approach in human health and disease. Proteomics and transcriptomics techniques are also required to fully elucidate the host-microbiome interaction (Kinross, Darzi and Nicholson, 2011).

4.4.3 | Impact on Saliva-Derived Biomarkers for Disease

Due to the ease of collecting saliva, and how it can reflect most of the compounds found in blood, it is considered to be a great promise in the identification of novel biomarkers for disease. As population-based screening programmes usually require a single sampling time point to measure an individual or panel of biomarkers, saliva-derived biomarkers will only be of use if their levels do not fluctuate over time. In this portion of work, the salivary metabolome has been shown to be stable over a one year period, based on a large sample size of 40 participants giving a total of seven samples each. This finding suggests that the salivary metabolome is sufficiently stable to make saliva-derived disease biomarkers plausible. Additionally, the majority of studies of the salivary metabolome have focused on the use of NMR spectroscopy, which has a lower sensitivity than mass spectrometry based approaches. Therefore, this portion of work has suggested that low abundant metabolites, which may not be detected using NMR spectroscopy, are also stable over a one year time period (Mamas *et al.*, 2011).

4.5 | Conclusions and Future Work

This portion of work aimed to characterise the variability of the human salivary microbiome and metabolome over time, namely a one year period with sampling every two months. With the fields of microbiomics and metabolomics showing promise in helping to understand health and disease, and to provide biomarkers for a range of human disorders, understanding their temporal variability is fundamental in establishing their usefulness and validity. To this end, this portion of work has shown that the human salivary microbiome displays temporal stability in terms of bacterial diversity but not load, over a one year period. The salivary metabolome showed a similar degree of temporal stability. A significant positive relationship between pH and the α-diversity of the microbiome in a sub-group of ten participants was also observed, which may have implications in oral diseases associated with increased salivary acidity.

Saliva was chosen as the biofluid in this portion of work because it is non-invasive to collect and shows promise as a source of novel biomarkers for disease. Although the salivary microbiome, in terms of bacterial diversity, and metabolome have been shown to be temporally stable, this finding cannot be extended to the microbiome and metabolome found in other human systems, such as the gastrointestinal tract. Furthermore, the participants in this portion of work may not be accurate representatives of the microbiome and metabolome found under disease conditions. It may be disease-related biomarkers, in both the microbiome and metabolome, are under a greater degree of instability than that found in healthy individuals.

Although the salivary microbiome, in terms of bacterial diversity, and metabolome have been shown to be temporally stable using the methods employed in this portion of work. As previously discussed, the use of metagenomic sequencing may reveal species-level taxonomic or functional capability changes that are not visible using 16S rRNA amplicon sequencing. Furthermore, metabolomics can only reveal changes in relatively low molecular weight compounds, usually those below 1.5 kDa, and it may be that temporal instability exists in the proteome or lipidome of human saliva.

CHAPTER 5 | Humans and Their Hidden Companions Cross Antarctica

CHAPTER SUMMARY | The effects that prolonged human space travel could have on the human microbiome and metabolome are unclear, and could potentially impact the successful outcome of manned space travel, such as to Mars. In this portion of work, the expedition members of the Trans-Antarctic Winter Traverse (TAWT) gave stool, saliva, and blood plasma samples for each of the eight months of travel, in addition to a preliminary baseline sample. The bacterial diversity of the salivary and stool microbiomes was determined through amplicon sequencing of the V3 to V4 region of the 16S rRNA region, and bacterial load through quantitative PCR. Metabolome profiles of raw saliva, saliva supernatant, stool, and blood plasma were constructed using negative mode LTQ-MS. Additionally, stool water content and the pH of saliva supernatant, raw saliva and blood plasma was determined. Significant ($P < 0.001$) differences were seen between baseline salivary bacterial load and all other time points, but not between participants. Additionally, bacterial diversity was significantly ($P = 0.002$) lower in Baseline samples than all other sampling months, though individual differences were also significant ($P = 0.016$). At the taxonomic level of classification, significant differences in the Bacteroidetes ($P = 0.008$) phylum, and the *Desulfotomaculum* ($P = 0.013$), *Veillonella* ($P = 0.016$), *Prevotella* ($P = 0.022$), *Megasphaera* ($P = 0.041$), *Bacillus* ($P = 0.047$), and *Rothia* ($P = 0.047$) genera were evident between sampling months. The stool microbiome showed no significant differences in bacterial load between participants ($P = 0.131$) or sampling month ($P = 0.867$), but bacterial diversity was significantly ($P < 0.001$) different between participants. Stool water content ($P < 0.001$), pH of raw saliva ($P < 0.001$), and saliva supernatant ($P = 0.001$) all showed significant differences between participants. No differences were evident in the metabolomes of any of the four biofluids profiled by either participant or sampling month. Overall, the human salivary microbiome shows significant alterations in bacterial load and diversity as a result of the TAWT expedition. However, the stool microbiome and metabolomes of all four biofluids studied do not. Therefore, the human microbiome and metabolome is differentially affected based on sample site, which could have important implications for future space travel.

5.1 | Introduction

The human microbiome and metabolome have both been shown to be an important component of health and disease. A number of stresses that involve deviations from a person's normal lifestyle can lead to changes in the human microbiome and metabolome. Long-term exploration, such as that envisioned for future human space travel, such as manned travel to Mars, could have significant implications for the human microbiome and metabolome. This could impact upon the health of participants, and the potential success of expeditions.

Due to the nature of human space travel, analogous terrestrial expeditions to the isolation and environmental and physiological stress that may be experienced during space travel have to be held to gauge the effects that they may have on the human microbiome and metabolome. Concurrent to the TAWT expedition, the White Mars scientific project aimed to utilise the unique stressors that expedition members experienced to explore how the human body is affected.

5.1.1 | The Human Stress Response

Stressful stimuli, including physiological, environmental and psychological, and the corresponding stress response are central to the maintenance of homeostasis. However, the stress response can also lead to adverse consequences in humans, with a number of diseases including obesity, heart disease, and depression, having stress as a risk factor (Groeneweg *et al.*, 2011).

The term stress can refer to a range of conditions and responses, where the stimulus can range from mild to severe, which lead to a host response, such as the release of hormones. However, an emerging concept in the field of stress research is that the term stress should only refer to a situation in which the demand of a stimuli exceeds the natural regulatory ability of the organism to maintain homeostasis, and in particular where these stimuli are unpredictable and uncontrollable (Koolhaas *et al.*, 2011).

In response to stressful stimuli, the human body employs physiological and behavioural systems, including the central nervous system and the peripheral adaptive response. The human stress response is predominately a hormonal one, relying on the endocrine system to produce hormones to maintain homeostasis and adapt to the stressful stimuli (Groeneweg *et al.*, 2011). Stress is also able to increase the permeability of the gut, Figure 5.1, leading to increased levels of pro-inflammatory cytokines, such as interferon gamma (IFNγ) and interleukin six (IL-6). In turn, these can alter activity of indoleamine 2,3-dioxygenase activity in the liver, lowering tryptophan availability to the brain and altering levels of 5-hydroxytryptamine, also known as serotonin. Altered serotonin levels can lead to increased levels of corticotrophin-releasing factor (CRF) and arginine vasopressin in the adrenal cortex, resulting in higher levels of cortisol (Dinan and Cryan, 2012).

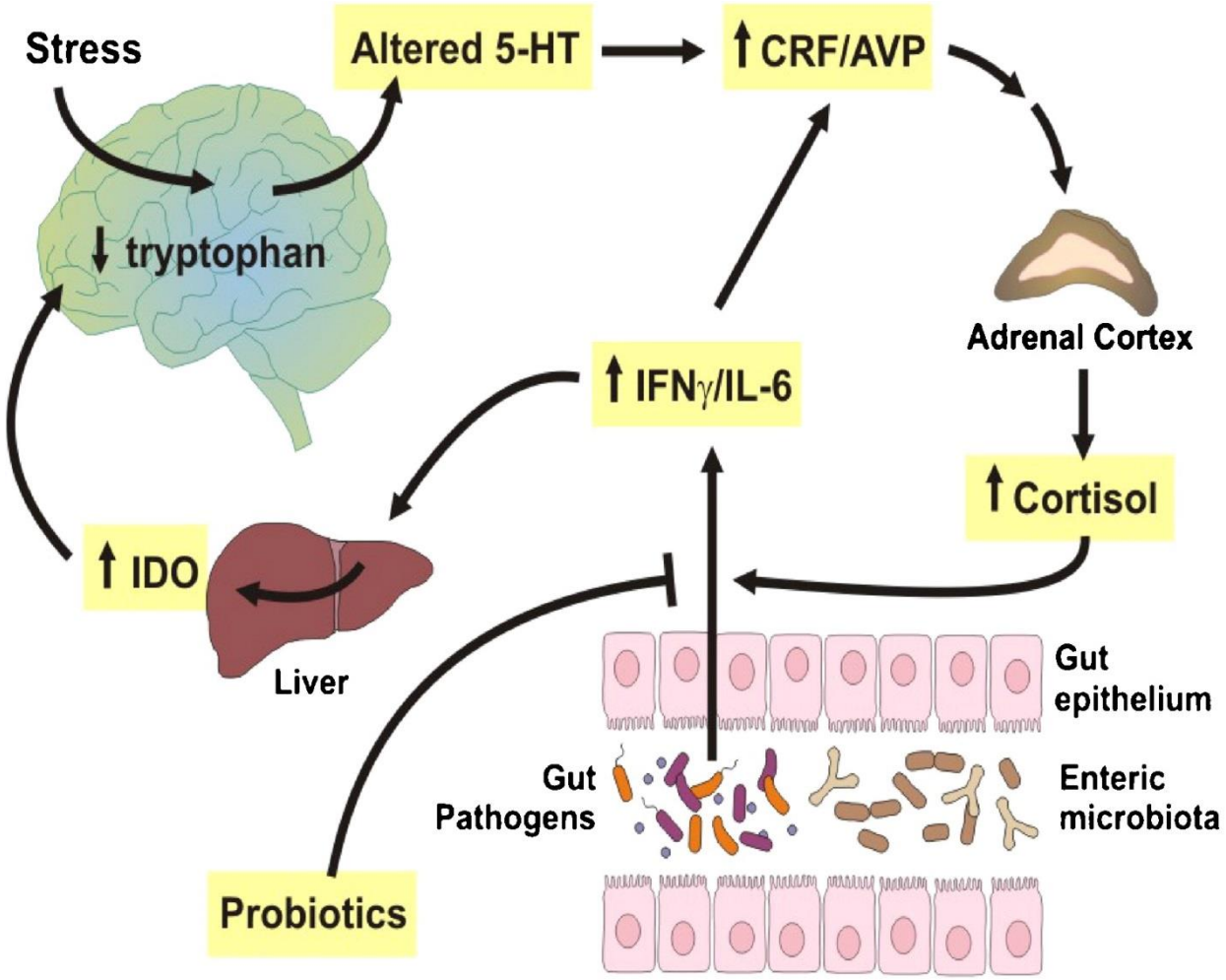


FIGURE 5.1 | Stress Effects on Human Gut Permeability
Exposure to environmental, physiological, or psychological stressors can lead to increased human gut permeability, allowing bacteria and bacterial antigens to cross the gut epithelial layer. This can result in an immune response leading to increased pro-inflammatory cytokines, affecting a number of systems, including the brain; altering hormonal levels. Certain probiotic bacteria have been shown to alter the gut permeability following stress. Figure taken from Dinan and Cryan, (2012).

Alterations in the gut-brain axis have been linked to a range of psychological and physiological disorders, including depression and inflammatory bowel disease (Konturek, Brzozowski and Konturek, 2011). Additionally, there is a high incidence of co-morbidity between stress-related psychiatric symptoms, such as anxiety, with gastrointestinal disorders. However, to date, the exact mechanisms involved in the gut-brain axis, and the bidirectional communication between the gut and brain are still to be elucidated. Ongoing investigations into elucidating these mechanisms include assessing the impact of probiotics, antibiotics, and pathogen-induced dysbiosis on the gut-brain axis (Cryan and O'Mahony, 2011).

5.1.2 | Space Travel and the Human Microbiome and Metabolome

The human microbiome has been shown to be an essential component in the maintenance of homeostasis. Furthermore, it has been shown to modulate the stress response, either through the production or alteration of metabolites. To date, the field of medicine interested in the effect of prolonged space travel has focused on improving the provision of artificial life-support systems, and the human microbiome has received only minimal attention (Saei and Barzegari, 2012). This may be important as the stress of microgravity has been shown to alter host-microbiome interactions and even increase the virulence of a number of human pathogens. Microgravity could alter the human microbiome to an extent that overall dysbiosis gives rise to a number of diseases that are not associated with a single microbial agent (Foster, Wheeler and Pamphile, 2014).

The environment that astronauts would experience during prolonged space travel offers a number of unique characteristics. For example, the environmental microbiome found on the International Space Station appears to consist of a large number of human-associated microbes, with around 90% of pyrosequenced reads associated with Actinobacteria. Interestingly, members of this phylum have been associated with structural damage to buildings, and to allergic-respiratory conditions which may impact upon inhabitants (Venkateswaran *et al.*, 2014). As well as posing unique environmental conditions, prolonged space travel has inherent environmental stressors, such as microgravity, but also in exposure to radiation, which is considered one of the primary barriers to a manned mission to Mars. Work with

mice exposed to whole-body radiation, with samples taken after exposure to a LD_{50} and LD_{30} dose showed that the gut microbiome shifts significantly. Of particular note was the shift in the gut microbiome after irradiation to one which included increased levels of Enterobacteriaceae, which contains a number of known pathogens (Karouia *et al.*, 2014).

Space travel itself is not the sole means of investigating how prolonged space travel could affect the human microbiome. Recent isolation studies that attempt to mimic the isolation and physiological and psychological effects of prolonged space travel have measured changes in the structure and function of the human gut microbiome. In the Mars-500 study, changes in the taxonomic composition of the gut microbiome were seen within the first 14 days, but no such changes were seen in its core functional capacity, with a reversion to the initial microbiome composition beginning to be seen within two weeks of the study ending. Although slight taxonomic changes were observed in the microbiome, no adverse effects were consequently seen in the host, suggesting that the microbiome is highly adaptable to the environmental conditions and stressors that may be involved in prolonged space travel (Mardanov *et al.*, 2013).

The human metabolome reflects changes in the body's metabolism, and how it has altered to adapt to stressors. As an analogy to prolonged space travel, the environmental and physiological conditions presented by the Antarctic have previously been shown to alter the blood plasma metabolome, reflecting changes associated with altered liver and kidney function (Yadav *et al.*, 2014).

Radiation is considered to be one of the biggest environmental stressors facing those attempting prolonged space flight (Chancellor, Scott and Sutton, 2014). As with the human microbiome, the metabolome has also been shown to be significantly affected by radiation. Although limited to animal studies, long-term exposure to radiation has been shown to alter the intestinal metabolome of mice, with nucleotide and amino acid metabolism being significantly affected. Interestingly, the type of

radiation exposure was also shown to have preferential effects on the gut metabolome, with ^{56}Fe radiation leading to particularly altered dipeptide metabolism (Cheema *et al.*, 2014).

The use of metabolomics to investigate the human metabolome and how it is affected by prolonged space travel, has given rise to the prospect of personalised medical approaches to counter the negative consequence of the induced stress response. Metabolomics, in combination with other 'omic

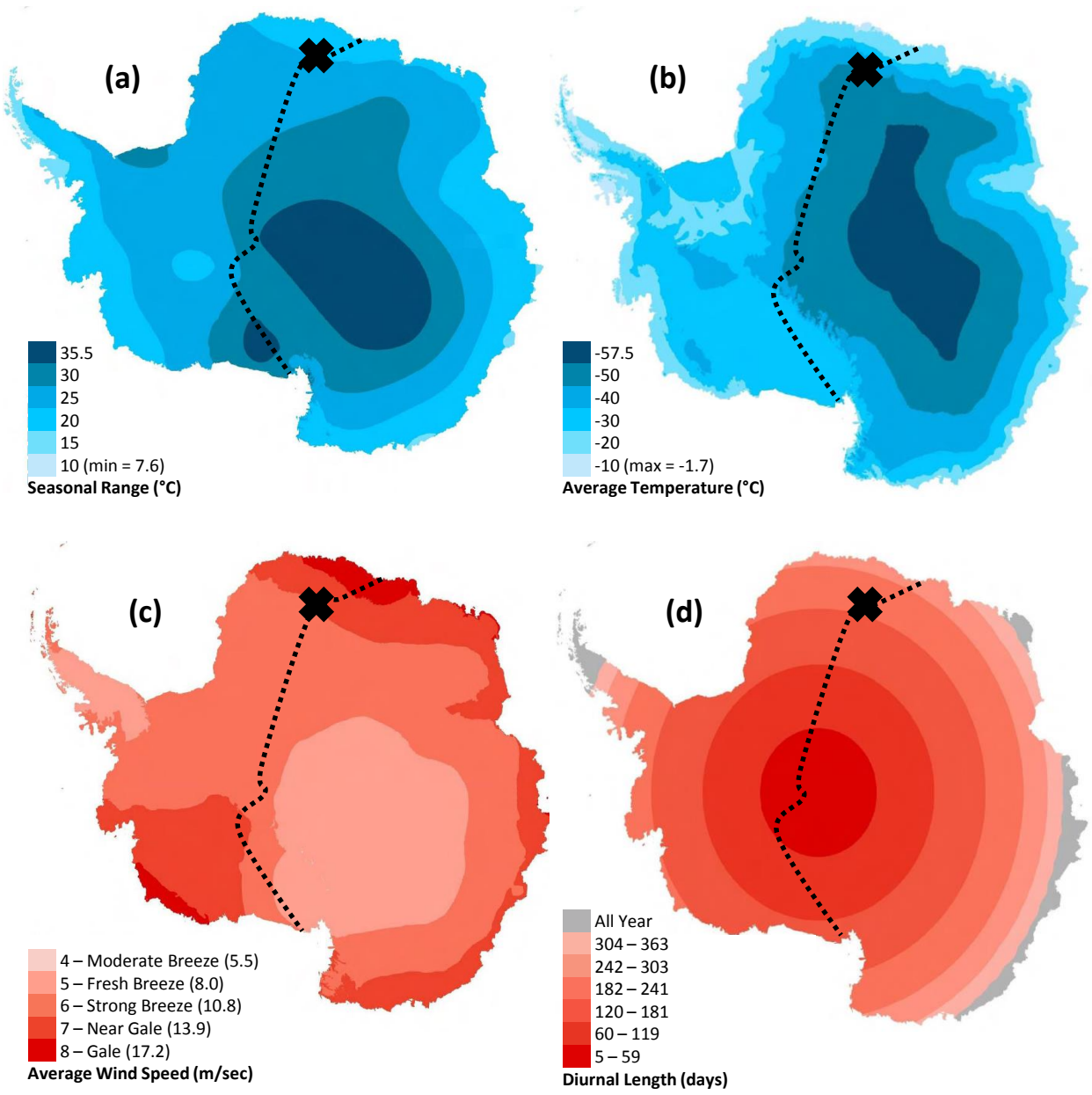


FIGURE 5.2 | Antarctic Conditions Facing TAWT Expedition
The Antarctic poses a highly inhospitable environment, in terms of (a and b) temperature, (c) wind speed, and (d) lack of sun light. Black dotted line displays the approximate planned route of the TAWT expedition before abandonment part-way through. X displays approximate location of stationary phase of TAET expedition following traverse abandonment. Figures adapted from Morgan *et al.*, (2007).

technologies such as proteomics and genomics, is able to assess how individuals respond differently to the same level of environmental or physiological stressor, such as radiation. Additionally, metabolomics could be used to tailor individualised countermeasures to a range of stressors that could be used to reduce the negative effects of prolonged human space travel (Schmidt and Goodwin, 2013).

5.1.3 | The Trans-Antarctic Winter Traverse

Considered to be the last major polar challenge, the Trans-Antarctic Winter Traverse expedition aimed to travel the 2,000 miles required to cross the Antarctic in the winter (March to September), from Crown Bay to the Amundsen-Scott South Pole Station, and then to Ross Island. Environmental conditions in the Antarctic include temperatures reaching down to -90°C and near complete darkness, Figure 5.2. The five man team travelled predominately on the polar plateau, which averages 2,160 m in thickness. In addition to the inhospitable environmental conditions, the terrain can also be dangerous with crevasses posing particular threat.

The initial plan for the TAWT expedition was to complete the Antarctic crossing from March 2013 to September 2013, with several months either side to prepare for the expedition and subsequent uplift. However, in June 2013 after travelling approximately 300 km, and climbing from sea level to an altitude of 3,000 m, the TAWT expedition team decided to abandon the rest of the distance. This was because of a range of technical and logistical issues, as well as the danger posed by an unexpected crevasse field ranging up to 100 km in the proposed distance of travel. Nevertheless, the environmental and physiological conditions that the five man expedition team were in for the remainder of the time period in Antarctica still provide for a rare analogous situation to prolonged space travel.

5.1.4 | Aims and Objectives of Chapter

The prospect of prolonged human space travel poses a number of novel questions for the role of the human microbiome and metabolome in health and disease. From laboratory *in vitro* and preliminary,

pilot-stage *in vivo* studies, the human microbiome and metabolome have been suggested to be important in maintaining the health of participants in prolonged human space travel. However, with prolonged human space travel not currently being envisaged, terrestrial-based analogies to the environmental and physiological stresses which may be encountered may help to elucidate the response of the human body to them, and reveal important information that may aid in the planning and implementation of prolonged space travel. To this end, the participants of the Trans-Antarctic Winter Traverse expedition conducted during 2013/2014 will provide samples and meta-data to help explain:

1) The effects of physiological and environmental stress on the human salivary and stool microbiome.

2) The effects of physiological and environmental stress on the human salivary, stool, and blood plasma metabolome.

5.2 | Materials and Methods

The White Mars project, from which samples were received for this study, formed part of the TAWT expedition, which took place from December 2012 to September 2013. The White Mars project was hosted by King's College London's Centre of Human and Aerospace Physiological Sciences, who gained the necessary ethical approval to collect samples. This component of the White Mars project, that forms the contents of this Chapter, received additional ethical approval from the Aberystwyth University Research Ethics Committee. Informed consent was obtained from all study participants, and all participant information was link anonymised by the expedition doctor.

5.2.1 | Participant Recruitment and Sampling

A total of five participants took part in the White Mars element of the TAWT. All sampling was completed by the expedition doctor, and samples stored at ≈ -40°C throughout the expedition. For stool and saliva, a baseline sample was taken from all participants prior to the start of the expedition and stored at -30°C until all samples had returned from the TAWT expedition. During the expedition, stool, saliva and plasma samples were taken at monthly intervals, over an eight month period, alongside individual physiological information and measurements. Samples were stored at King's College London at -30°C for approximately two weeks, until transferred to Aberystwyth laboratories on dry ice, over a six hour period.

Throughout the TAWT expedition, all participants consumed a similar high-calorie diet of freeze-dried food and consumed similar liquid intake. At night, all participants sheltered within the expedition transport caboose.

5.2.2 | Sample Processing

Upon arrival at Aberystwyth laboratories, the 4 mL raw saliva samples were thawed at 4°C, vortexed to homogenise mixture, and separated into 2 mL aliquots in sterile 2 mL microcentrifuge tubes. One

aliquot was immediately frozen at -80°C and treated as raw saliva, and the remaining aliquot underwent centrifugation at 11,000 x g for ten minutes at 4°C. The resulting supernatant was removed and transferred to a sterile 2 mL microcentrifuge tube and immediately frozen at -80°C for approximately one month until analysed using LTQ-MS. The remaining pellet was also immediately frozen at -80°C until used for DNA extraction, which was completed within seven days receipt at Aberystwyth laboratories.

Blood samples were thawed at 4°C and underwent centrifugation at 11,000 x g for ten minutes at 4°C to ensure no separation (indicative of mixed blood) of sample, and 2 mL transferred to a sterile 2 mL microcentrifuge tube and immediately frozen at -80°C. Stool samples were transferred to sterile, pre-weighed glass vials, and weighed to determine wet weight. Stool samples were frozen at -20°C and transferred to a freeze-drier set at -50°C. Stool samples were then weighed every 24 hours until they had reached a stable weight, determined as a weight reduction, over a 24 hour period, of less than 5%. After freeze-drying, 100 mg of each stool sample was transferred to a sterile 2 mL microcentrifuge tubes and stored at -80°C, with the remaining stool sample transferred to a separate sterile 2 mL microcentrifuge tube and stored at -80°C. Total genomic DNA extraction was completed on the 100 mg aliquot within seven days of receipt at Aberystwyth laboratories, and the remaining sample analysed using LTQ-MS within one month.

5.2.3 | Total Genomic DNA Extraction and Purification

For saliva samples, DNA was extracted from the saliva pellet using a FastDNA SPIN kit for Soil (MP Biomedical). DNA extraction was completed following manufacturer's instructions, with bead beating carried out in a FastPrep-24 machine (MP Biomedical) with three cycles at speed setting 6.0 for 30 seconds, with on ice for 60 seconds between cycles. Genomic DNA was eluted with 50 µL of DES and dsDNA concentration determined, in duplicate, using 2 µL on the Epoch spectrometer system (BioTek). For stool samples, DNA was extracted from 100 mg of freeze-dried stool using a FastDNA SPIN kit for Soil (MP Biomedical), as previously described, except that an additional wash step was completed, and genomic DNA was eluted with 100 µL of DES. DNA extracts from stool underwent an additional

purification step to remove humic acid, which can be a common contaminant that inhibits downstream PCR. To 100 µL of extracted DNA from stool, 10 µL of a 3 M (pH 5.2) solution of sodium acetate was added. After vortexing for ten seconds, 275 µL of ice-cold, HPLC grade ethanol was added, and the mixture cooled on ice for 30 minutes. The cooled mixture then underwent centrifugation at 11,000 x g for 15 minutes at 4°C and the supernatant removed. An additional 1 mL of ice-cold, HPLC grade ethanol was added and the sample placed on a vortex mixer for ten seconds before undergoing centrifugation at 11,000 x g for 15 minutes at 4°C. The samples were then air dried in a sterile laminar flow hood for ten minutes, after which the DNA was reconstituted in 100 µL of DES buffer. DNA concentration was determined, in duplicate, using 2 µL on the Epoch spectrometer system (BioTek). Due to variations in DNA loss caused by this additional purification step, samples were concentrated in a DNA 120 Speed Vac (Savant Instruments, New York, USA) through complete removal of DES buffer. Samples were then reconstituted in DES buffer to reach the same DNA concentration as after initial DNA extraction.

5.2.4 | 16S rRNA Quantitative PCR

Quantitative PCR was carried out, in duplicate, on neat extracted DNA against standards created by amplifying the 16S rRNA gene of the five baseline samples, separately for stool and saliva samples. This was completed through amplification of the 16S rRNA gene in a 20 µL reaction volume consisting of 10 µL of 2 x BioMix (BioLine), 0.25 µL each of 27f (5'-AGA GTT TGA TCC TGG CTC AG-3') and 1389r (5'-ACG GGC GGT GTG TAC AAG-3') primers (Hongoh, Ohkuma and Kudo, 2003) to give a final concentration of 500 nM, 1 µL of neat extracted DNA, and 9.5 µL of PCR Grade Water (Roche). The reaction volumes were then subjected to PCR consisting of 94°C for two minutes, 30 cycles of 94°C for 45 seconds, 55°C for 45 seconds, and 72°C for 90 seconds, followed by a final elongation step of 72°C for seven minutes. The resulting PCR products were combined and purified using an Isolate II PCR and Gel Extraction purification kit (BioLine), following manufacturer's instructions, and quantified using an Epoch spectrometer as previously described. The resulting DNA concentration was used to estimate the total number of 16S rRNA gene copies and serial dilutions of 10^{10}, 10^{8}, 10^{6}, 10^{4}, 10^{2}, and 10^{0} made.

Quantitative PCR was completed on neat extracted DNA against standards with each reaction completed in 25 µL volumes, each consisting of 12.5 µL 2 x SYBR Green Mastermix (Life Technologies), 0.25 µL of each EubF1 (5'-GTG STG CAY GGY TGT CGT CA-3') and EubR1 (5'-ACG TCR TCC MCA CCT TCC TC-3') primer (Maeda *et al.*, 2003), in a final concentration of 400 nM, 11 µL of PCR Grade Water (Roche) and 1 µL of neat DNA extract. Reactions were run using a C100 thermal cycler (BioRad, Hercules, USA) and CFX96 optical detector (BioRad), with data captured using CFX Manager software (BioRad), under conditions of 95°C for ten minutes, 40 cycles of 95°C for 15 seconds and 60°C for 60 seconds, followed by a melt curve consisting of a temperature gradient of 60°C to 95°C in 0.5°C increments, each for five seconds. Where shown, *P* values represent the significance of one-way ANOVA tests.

5.2.5 | 16S rRNA Amplicon Preparation

Taxonomic identification of bacterial component of the microbiome was carried out through sequencing of the V3 to V4 regions of the 16S rRNA gene on the Illumina MiSeq platform. Firstly, the V3 to V4 region of the 16S rRNA gene was amplified through duplicate PCR with locus specific primers, alongside negative water controls. In a 25 µL reaction volume, 12.5 ng of extracted DNA, or 2.5 µL of PCR grade water for negative controls, was added to 12.5 µL of 2 x Accuzyme Mix (BioLine) and 5 µL each of a 1 µM concentration of 319f primer (5'– CCT ACG GGN GGC WGC AG–3') with Illumina forward overhang adapter sequence (5' – TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG-3') and 806r primer (5'–GAC TAC HVG GGT ATC TAA TCC–3') with Illumina reverse overhang adapter sequence (5'- GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G-3') as detailed by Klindworth *et al.*, (2013). The reaction mix underwent PCR consisting of 95°C for three minutes, followed by 25 cycles each of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by a final elongation step of 72°C for five minutes. Each duplicate PCR volume was confirmed through visualisation on a 2% agarose gel, after being run for 120 minutes at 100 volts (≤ 80 mA) in 1% TAE buffer. Following confirmation of PCR success, corresponding reaction volumes were combined and purified using an Isolate II PCR and Gel Extraction kit (BioLine), following manufacturer's instructions, with elution into 20 µL of kit buffer. Following purification, a second PCR was completed to attach Illumina adaptors to amplified products to

allow for multiplexed amplicon sequencing on the Illumina MiSeq platform. To make a final reaction volume of 25 µL, 2.5 µL of purified PCR product from the previous step was added to 12.5 µL of 2 x Accuzyme Mix (BioLine), 5 µL of PCR Grade Water (Roche) and 2.5 µL each of the relevant Nextera XT Index Primer 1 (N7##) and Nextera XT Index Primer 2 (S5##) (Illumina) as detailed in Chapter 5 Appendix, Supplementary Table 5.1. The reaction mix underwent a limited cycle PCR consisting of 95°C for three minutes, eight cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by a final elongation step of 72°C for five minutes. To remove non-combined adaptors, the entire reaction volume was run on a 2% agarose gel, in TAE buffer, for 120 minutes at 100 volts (≤ 80 mA). The gel was visualised using a DR195M Transilluminator (Clare Chemical Research, Colorado, USA) and each PCR product excised using a sterile scalpel blade. PCR products were purified using an Isolate II PCR and Gel Extraction kit (BioLine), following manufacturer's instructions, with elution into 20 µL of kit buffer. Purified PCR products were quantified using a Quant-iT dsDNA High Sensitivity assay kit and a Qubit fluorometer (Life Technologies).

5.2.6 | 16S rRNA Amplicon Sequencing and Analysis

Individual sample libraries were pooled together in equimolar concentration and sequenced, along with 20% PhiX DNA as a control for low diversity, on the Illumina MiSeq platform using MiSeq v3 reagents for a 2 x 300 bp run at the IBERS Translational Genomics Facility, Aberystwyth University. After sequencing, sample reads were demultiplexed and trimmed to remove primer sequences. Overlapping paired-end reads were merged using the MG-RAST metagenomic analysis pipeline, dereplicated and trimmed for quality using default parameters (Meyer *et al.*, 2008). Taxonomic alignments of sequences was completed using the Ribosomal Database Project (Cole *et al.*, 2009) facility, with only those sequences with a minimum alignment identity of 97%, maximum e-value of 1×10^{-5}, and a minimum alignment cut-off of 15 being used. Sequences were analysed using the principal component analysis, α-diversity measurement using the Shannon diversity index, and taxonomic assignment facility within MG-RAST, and exported into Microsoft Excel 2010 and MINITAB 14 for additional analysis, with multivariate statistics completed using MetaboAnalyst 2.0 (Xia *et al.*, 2012). Sequence numbers for each sample were

normalised as a percentage composition of the total volume of sequences for each taxonomic level of classification. Where shown, *P* values represent the significance of one-way ANOVA tests. All sequence files are available under the MG-RAST project ID 11838: 'White Mars'. Raw sequence reads were deposited at the European Nucleotide Archive under primary accession number PRJEB9027 and secondary accession number ERP010081.

5.2.7 | LTQ-MS Metabolomic Fingerprinting

For raw saliva, saliva supernatant and blood plasma, samples were thawed at 4°C, and 200 µL transferred to a sterile 2 mL microcentrifuge tube, to which 30 mg of ≤ 106 µM acetone-washed glass beads (Sigma-Aldrich) were added. To this, 1520 µL of a solvent mix of HPLC grade methanol and chloroform, in a ratio of 4:1, was added. For stool samples, freeze-dried stool was reconstituted at a concentration of 100 mg/mL in a solvent mix of water, methanol and chloroform, in a ratio of 2:5:2 respectively. To homogenise the mix, samples were placed on a vortex mixer for five seconds and then milled for 30 seconds at 30 Hz. Following this, samples were shaken for 20 minutes at 4°C and then stored at - 20°C for 20 minutes to precipitate protein. Samples then underwent centrifugation at 11 000 x g for six minutes, at 4°C. For raw saliva, saliva supernatant and blood plasma samples, the resulting supernatant was removed and transferred to a sterile 2 mL microcentrifuge tube. For stool samples, the resulting supernatant was removed and diluted by 50% in the same water, methanol and chloroform solvent mix detailed earlier. From these, 70 µL of sample was transferred to an LTQ-MS vial and sealed. Samples were kept at -20°C until run, in a randomised order using an autosampler, with tray temperature kept constant at 15°C. For each sample, 20 µL was injected into a flow volume of 60 µL per minute water-methanol, in a ratio of 70% water and 30% methanol, using a Surveyor liquid chromatography system (Thermo Scientific, MA, USA). Data acquisition for each individual sample was conducted, in alternating positive and negative ionisation mode, over four scan ranges (15-110 m/z, 100-220 m/z, 210-510 m/z, 500-1200 m/z) on an LTQ linear ion trap (Thermo Electron Corporation, CA, USA), with an acquisition time of five minutes. Individual metabolite m/z values were normalised as a

percentage of the total ion count for each sample. Normalised abundances were subsequently analysed using MetaboAnalyst 2.0 (Xia *et al.*, 2012) and PyChem (Jarvis *et al.*, 2006).

5.2.8 | pH Measurements of Saliva and Plasma

Measurements of the pH of saliva supernatant was carried out using a B-212 Twin pH Meter (Horiba, Kyoto, Japan) after two point calibration using a pH 7 and pH 4 buffer. For pH measurements, 200 µL of saliva supernatant, after thawing at 4°C, was used; ensuring equal coverage of the two sensor points. After stabilisation of reading, pH value was recorded and the sensor washed with ultrapure water, and blotted dry. Data analysis was completed using Microsoft Excel 2010 and Minitab 14. Where shown, *P* values represent the significance of one-way ANOVA tests.

5.3 | Results

The TAWT expedition did not complete its initial intention to cross Antarctic during the winter (March to September). Due to the dangers posed by unexpectedly large crevasse fields, the expedition was halted in Month 4. This was preceded by three months of traversing, where a total distance of 300 km was travelled. After three months of remaining in their stationary position, the TAWT expedition took a further two months of traversing back to their original starting point.

The TAWT expedition team consisted of five males, with an average age of 37 (range 28 to 54). Table 5.1 gives each individual team member's physiological information, along with their anonymised participant identifier. Their weight, body mass index, and body fat percentage is given as a mean of their weekly measurements over the expedition period. Physiological information for each weekly measurement, for each participant, is given in Chapter 5 Appendix, Supplementary Table 5.2.

TABLE 5.1 | Participant Physiological Information
Individual participant information is detailed for each of the five members of the TAWT expedition. For weight, body mass index and body fat percentage (%), participant information is given as a mean of the weekly measurements taken with standard deviations shown italicised in brackets.

	A	**B**	**C**	**D**	**E**
Age	30	28	34	39	54
Gender	Male	Male	Male	Male	Male
Weight	77.0 (*3.0*)	80.2 (*2.4*)	81.1 (*1.2*)	64.3 (*0.7*)	65.8 (*2.0*)
Body Mass Index	26.8 (*1.0*)	27.8 (*1.0*)	25.6 (*0.4*)	21.4 (*0.2*)	22.4 (*0.6*)
Body Fat %	17.3 (*0.8*)	13.2 (*1.5*)	16.8 (*1.2*)	12.5 (*0.7*)	24.2 (*0.9*)

5.3.1 | Stool Water Content

Each monthly stool sample was freeze-dried until a stable weight was reached. This weight was then used to calculate the original water content of the stool sample. Figure 5.3 shows that stool water content maintains individual differences between participants over the nine monthly samples. Participant B was shown to have significantly ($P < 0.001$) lower stool water content than Participants A, D, and E and Participant C lower stool water content than Participant A only, Figure 5.3a. No significant ($P = 0.865$) differences were evident between sampling months.

5.3.2 | pH of Plasma and Saliva

The pH of saliva supernatant, raw saliva, and blood plasma was measured for each participant at each sampling month, Figure 5.4. No discernible pattern was evident for any of the three biofluids in regards to changes between sampling months. However, as with stool water content, individual differences were present between participants.Participant A was shown to have significantly ($P = 0.001$) lower saliva supernatant pH than Participants C and D. Additionally, Participant A's raw saliva pH was significantly ($P < 0.001$) lower than Participants C, D, and E. No such significant ($P = 0.054$) differences were evident

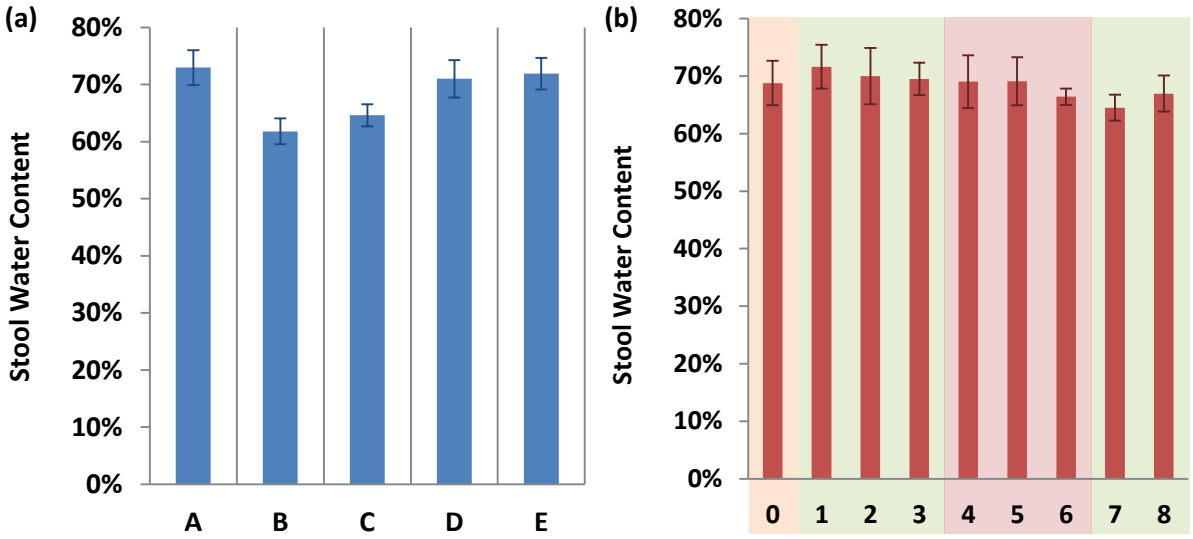


FIGURE 5.3 | Stool Water Content
Stool samples were freeze-dried and weighed every 24 hours until they reached a stable weight. Mean stool water content is shown by (a) participant and (b) sampling month. Significant ($P < 0.001$) differences were seen between participants, but not between sampling months. In (b), shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

between the blood plasma pH of participants. Additionally, no significant differences were evident in regards to sampling month for saliva supernatant pH (P = 0.134), raw saliva pH (P = 0.054) and blood plasma pH (P = 0.432).

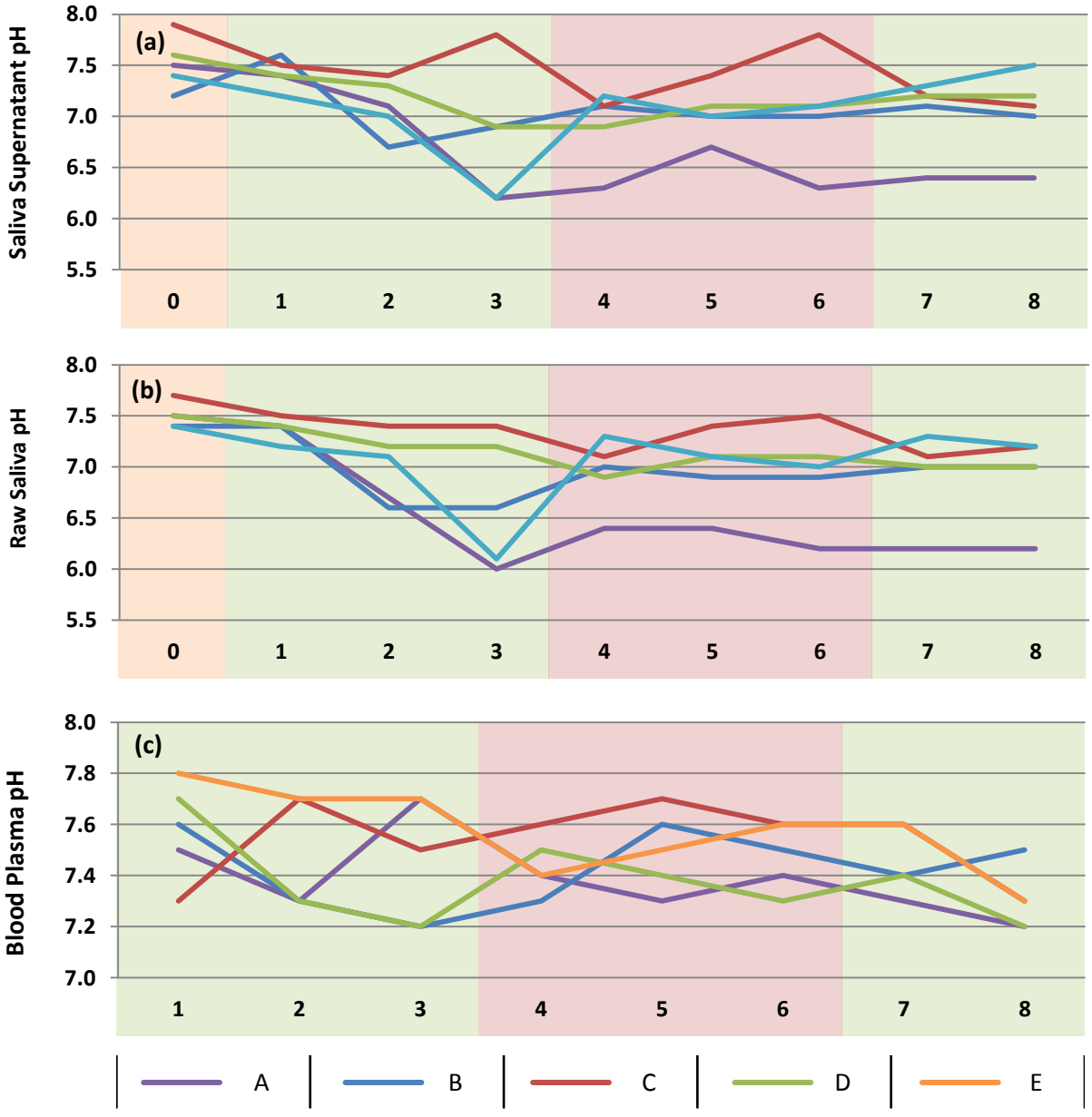


FIGURE 5.4 | Saliva Supernatant, Raw Saliva, and Blood Plasma pH Levels
Individual participant values for pH of (a) saliva supernatant, (b) raw saliva, and (c) blood plasma are shown for each sampling month. No discernible pattern is evident for any of the three biofluids in regards to sampling month, though Participant A had significantly (P = 0.001) lower saliva supernatant pH than Participants C and D, and significantly (P < 0.001) lower raw saliva pH than Participants C, D, and E. No such differences were observed between the blood plasma pH of participants. In all figures, shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods.

5.3.3 | Changes in Salivary Microbiome

To gauge changes in the salivary microbiome over the period of the TAWT expedition, both bacterial load and diversity was measured. The former was measured using quantitative PCR targeting the 16S rRNA gene, Figure 5.5. In regards to salivary bacterial load, no significant (P = 0.495) individual differences between participants, Figure 5.5a, was evident. However, significant (P < 0.001) differences were present between sampling months, with baseline samples showing lower bacterial load than at all other sampling points. Additionally, no relationship was shown to be present between salivary bacterial load and both saliva supernatant pH (R^2 = 6.4%, P = 0.093) and raw saliva pH (R^2 = 6.4%, P = 0.095).

Bacterial diversity in the salivary microbiome was measured through amplicon sequencing of the V3 to V4 regions of the 16S rRNA gene. Amplicon sequencing statistics for saliva are given in Chapter 5 Appendix, Supplementary Table 5.3, and show no significant differences between sequence number by participant (P = 0.326) or month (P = 0.792), or base pair number by participant (P = 0.385) or month (P = 0.748). However, significant differences were seen between sequence length between participants (P < 0.001) with Participant A having longer reads than all other participants by approximately four base

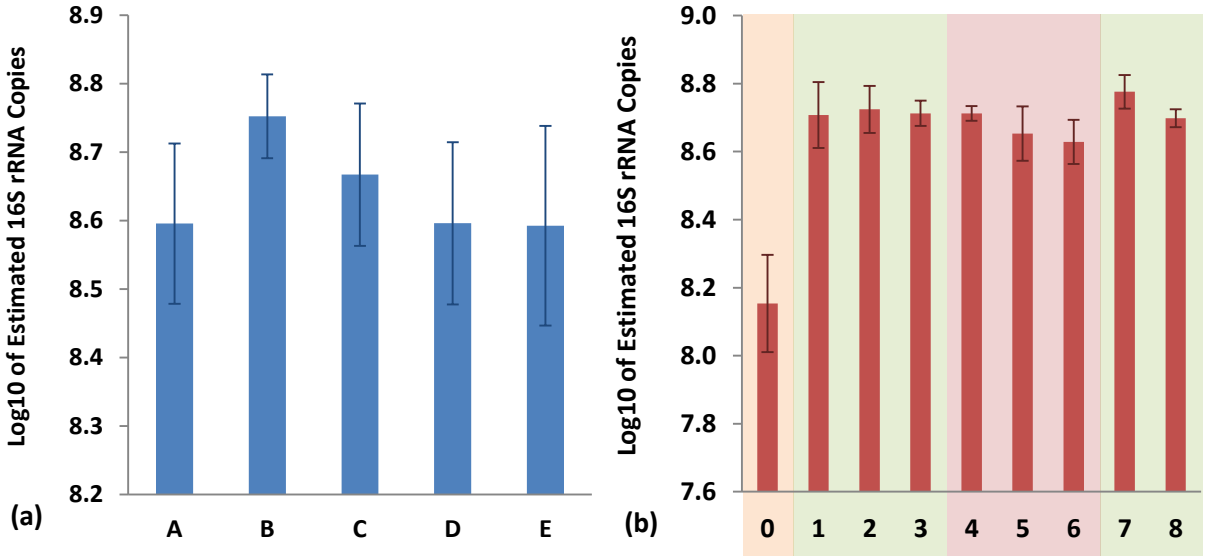


FIGURE 5.5 | Estimated Bacterial Load of Saliva
Means of the estimated bacterial load of saliva by (a) participant, and (b) sampling month are given. No significant (P = 0.495) differences were observed between participants. Significant (P < 0.001) differences were evident between sampling months, with the estimated bacterial load of saliva in the baseline sample lower than at all other sampling months. In (b), shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

pairs. No significant (P = 0.363) differences were evident between sequence length by sampling month.

Bacterial diversity within the salivary microbiome was measured through α-diversity calculated by the MG-RAST analysis pipeline, Figure 5.6. In regards to individual differences between participants, Participant A was shown to have significantly (P = 0.016) lower α-diversity than Participants B, D, and E, Figure 5.6a. Analysis by sampling months shows that, as with salivary bacterial load, α-diversity is significantly (P = 0.002) lower in baseline samples than at all other sampling months. No relationship was observed between α-diversity and both saliva supernatant pH (R^2 = 0.2%, P = 0.752) and raw saliva pH (R^2 = 0.0%, P = 0.900). A significant relationship was however, observed between α-diversity and salivary bacterial load (R^2 = 18.5%, P = 0.003).

Using the MG-RAST online platform, principal component analysis of the salivary microbiome, using classifications from the RDP database, was completed, Figure 5.7. No separation was evident between participants, Figure 5.7a, and sampling month, Figure 5.7b.

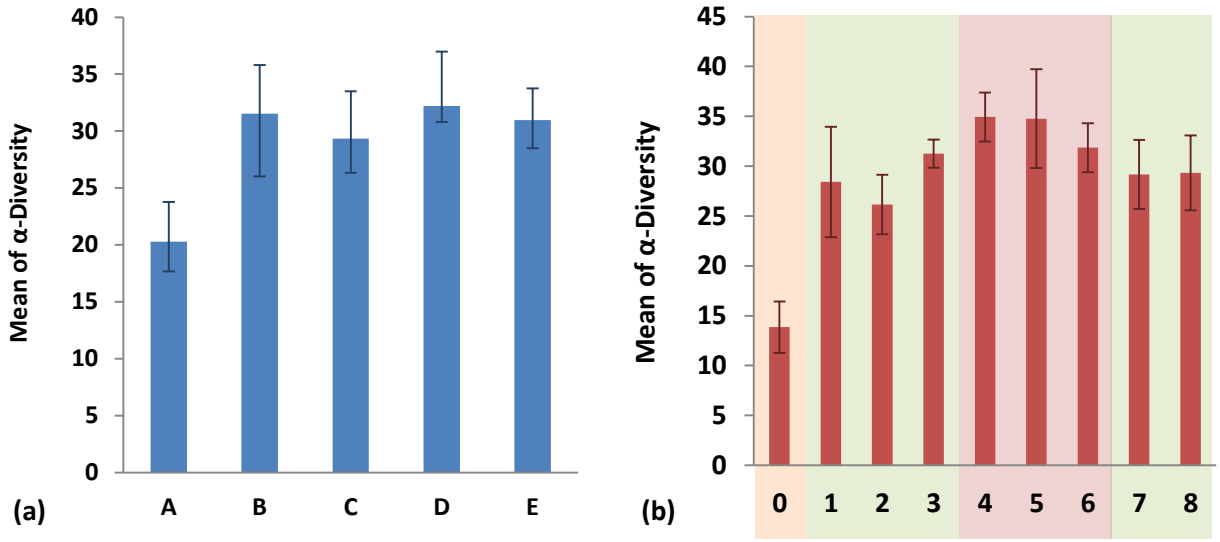


FIGURE 5.6 | α-Diversity of Salivary Microbiome
Means of α-diversity, calculated by MG-RAST analysis pipeline, of the salivary microbiome are shown by (a) participants, and (b) sampling month. Participant A was shown to have a significantly (P = 0.016) lower mean α-diversity than Participants B, D, and E. In regards to sampling month, the baseline samples were shown to have a significantly (P = 0.002) lower mean α-diversity than all other sampling months. In (b), shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

To this point, analysis of the salivary microbiome appears to suggest that changes between sampling months are important in determining its composition. To this end, phylum and genus-level taxonomic changes were analysed to identify those which significantly differed between sampling months. This revealed one phylum and six genera which were significantly different, Figure 5.8. At the phylum level of classification, Bacteroidetes was significantly (P = 0.008) lower in baseline samples than all other sampling months except Month 2. At the genus level of classification, *Desulfotomaculum* was significantly (P = 0.013) higher in Month 7 samples than in baseline and Month 1, 3, 5, and 8 samples. *Veillonella* was significantly (P = 0.016) lower in baseline samples than Month 1, 4, 5, 6, and 7 samples. *Prevotella* was significantly (P = 0.022) lower in baseline samples than in Month 1, 4, 5, 6, 7, and 8 samples. *Megasphaera* was significantly (P = 0.041) higher in Month 5 samples than in baseline and Month 1 samples. *Bacillus* was significantly (P = 0.047) higher in Month 4 samples than in baseline, Month 1, 3, and 7 samples. Finally, *Rothia* was significantly (P = 0.047) higher in baseline samples than in Month 1, 3, 4, and 7 samples. Of these phylum and genus-level taxonomic changes, no discernible pattern was seen in regards to activity during sampling month, namely traversing or stationary.

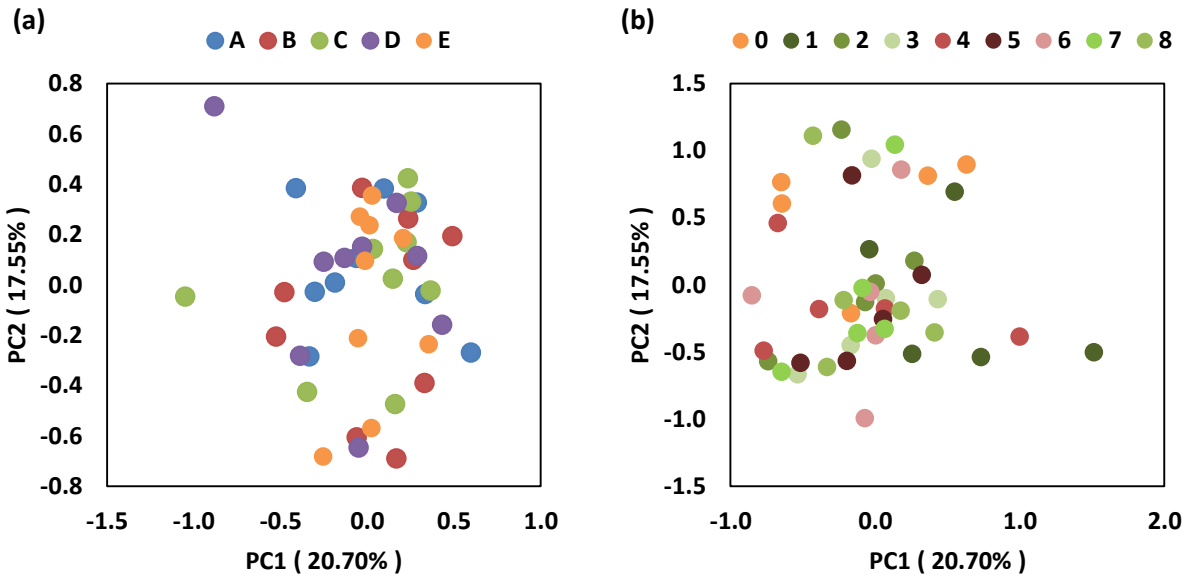


FIGURE 5.7 | Principal Component Analysis of Salivary Microbiome Composition
Using the MG-RAST online platform, principal component analysis was completed, using taxonomy classifications from the RDP database, at a 97% alignment, with other parameters at default. Resulting plots show no discernible separation by either (a) participant, or (b) sampling month. In (b), legend colour signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods.

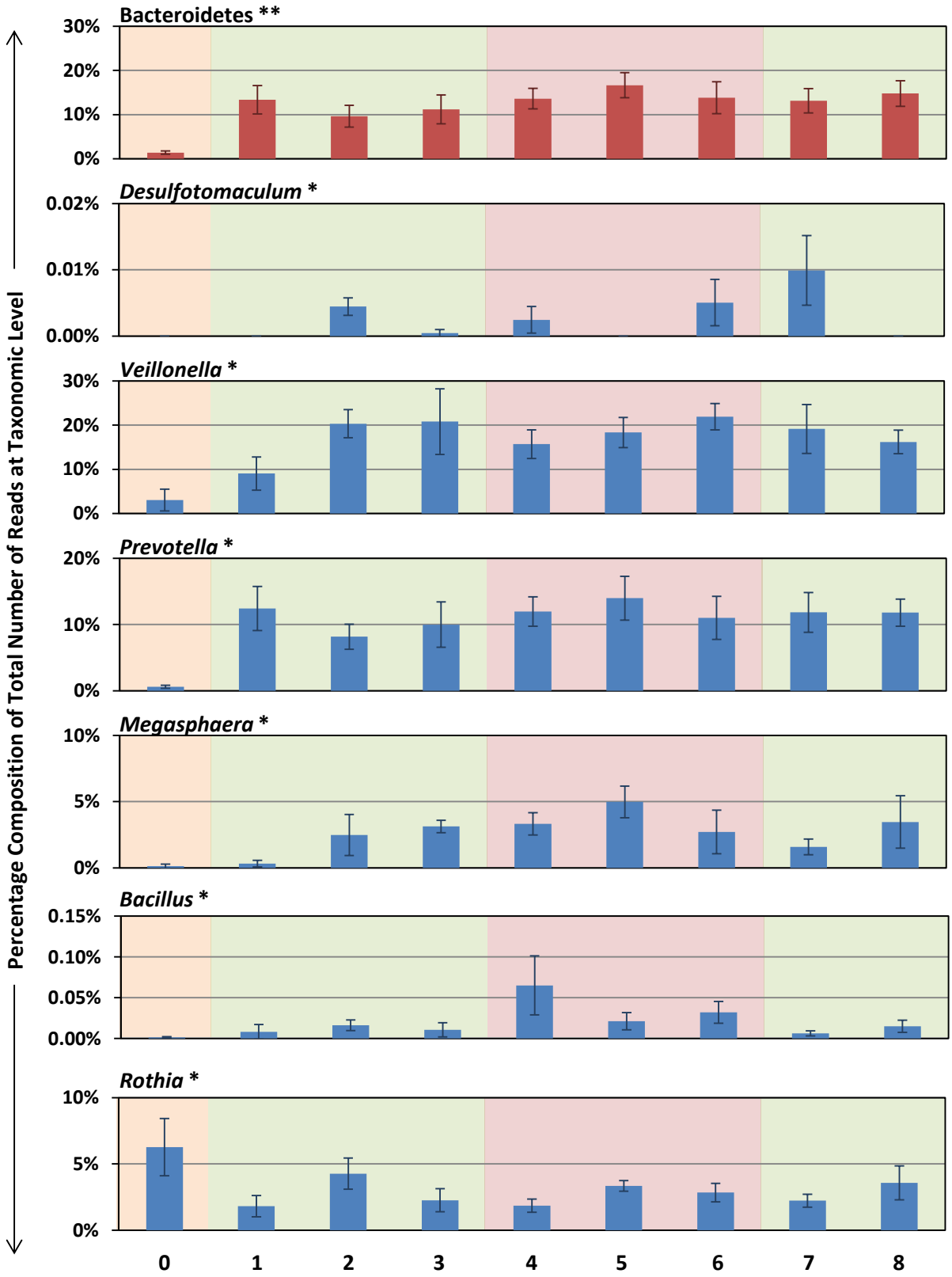


FIGURE 5.8 | Significant Phylum and Genus-Level Salivary Microbiome Changes by Month
Sampling month appears to have a significant impact on a number of salivary microbiome features. In regards to individual taxonomic changes, one phylum (red bars) and six genera (blue bars) showed significantly (* = $P < 0.05$, ** = $P < 0.01$) different levels of percentage composition, particularly in regards to baseline samples. In all figures, shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

5.3.4 | Changes in Stool Microbiome

As with the salivary microbiome, both bacterial load and diversity was measured. Bacterial load of stool was measured using quantitative PCR, Figure 5.9. In regards to individual changes between participants, no significant (P = 0.131) differences were observed, Figure 5.9a. Additionally, no significant (P = 0.867) differences were observed between sampling months, Figure 5.9b. No relationship was observed between the bacterial load of stool and stool water content (R^2 = 1.2%, P = 0.473).

Bacterial diversity in the stool microbiome was measured through amplicon sequencing of the V3 to V4 regions of the 16S rRNA gene. Amplicon sequencing statistics for stool are given in Chapter 5 Appendix, Supplementary Table 5.4. Significant (P = 0.003) differences were seen in regards to total base pair numbers, with Participant A having a higher total than Participants C and E.

No such significant (P = 0.946) differences were seen between sampling months. Additionally, significant (P = 0.003) differences were seen in total sequence number, with Participant A having a higher total than Participants C and E. No such significant (P = 0.946) differences were seen between sampling

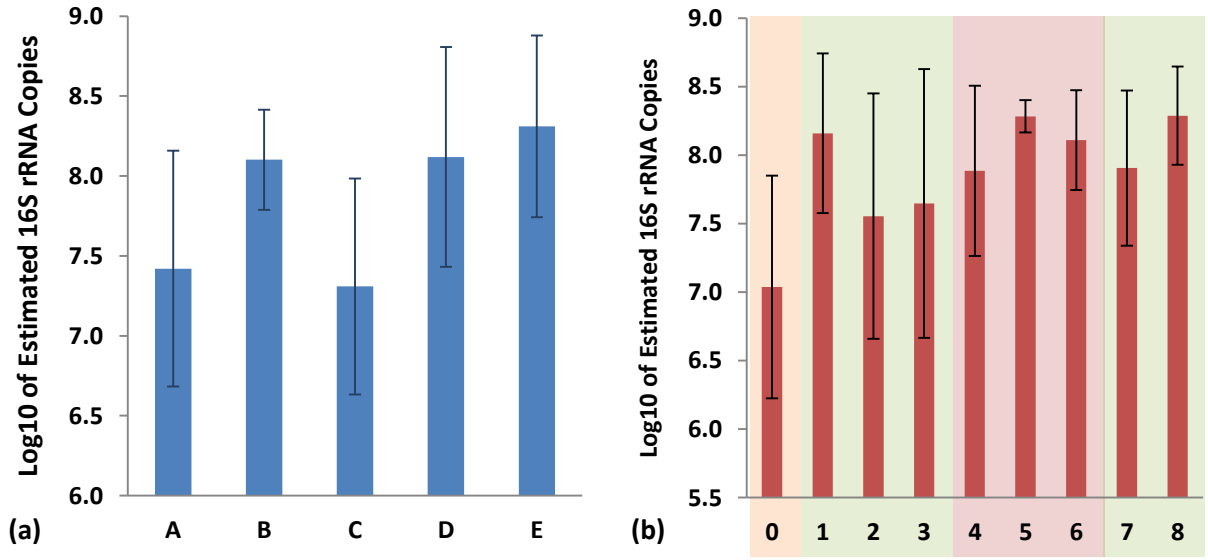


FIGURE 5.9 | Estimated Bacterial Load of Stool
Means of the estimated bacterial load of stool by (a) participant, and (b) sampling month are given. No significant (P = 0.131) differences were observed between participants. Additionally, no significant (P = 0.867) differences were evident between sampling months. In (b), shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

months. Furthermore, significant (P < 0.001) differences were seen in regards to average sequence length, with Participants A and B having shorter reads, by approximately one base pair, than Participants C, D, and E. No such significant (P = 0.993) differences were evident in regards to sampling month.

The bacterial diversity of stool was analysed through α-diversity values calculated using the MG-RAST analysis pipeline, Figure 5.10. In regards to individual differences, significant (P < 0.001) differences were evident, with Participant A showing a lower α-diversity than all other participants, and Participants B and D showing a lower α-diversity than Participants C and E, but higher than Participant A, Figure 5.10a. No such significant (P = 1.000) differences were evident in regards to sampling month, Figure 5.10b. No significant relationship was observed between the α-diversity of the stool microbiome and stool bacterial load (R^2 = 4.5%, P = 0.162) nor stool water content (R^2 = 6.1%, P = 0.102).

Using the MG-RAST online platform, principal component analysis of the composition of the stool microbiome was completed, Figure 5.11. Partial separation was seen in regards to participants, with Participants A and C showing the greatest degree of separation, Figure 5.11a. No separation was seen in

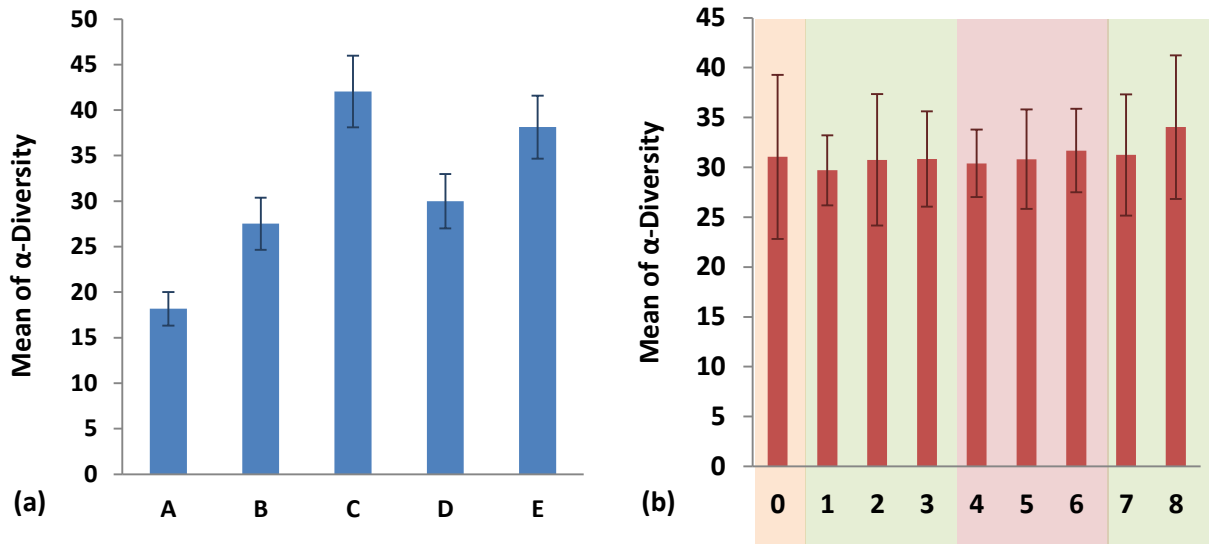


FIGURE 5.10 | α-Diversity of Stool Microbiome
Means of α-diversity, calculated by MG-RAST analysis pipeline, of the stool microbiome are shown by (a) participants, and (b) sampling month. Significant (P < 0.001) differences were observed between participants, with Participant A's α-diversity the lowest of all participants, and Participants C and E's α-diversity the highest. In regards to sampling month, no significant (P = 1.000) differences were observed. In (b), shading signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods. Error bars show one standard deviation around the mean.

regards to sampling month, Figure 5.11b.

Analysis of the stool microbiome, through principal component analysis and α-diversity values, appears to suggest that individual differences between participants are more substantial factors in determining the composition of the stool microbiome, than any changes associated with sampling month. Thus, phylum and genus-level changes between participants were analysed, Table 5.2, showing a total of four phyla and 38 genera significantly different in at least one participant.

5.3.5 | Stool, Plasma and Saliva Metabolome Changes

Using negative mode LTQ-MS, principal component analysis was completed on saliva supernatant and raw saliva, Figure 5.12, and stool and blood plasma, Figure 5.13, with both participant and sampling month groupings displayed. In all four biofluids, no substantial separation can be seen in regards to either participant or sampling month groupings.

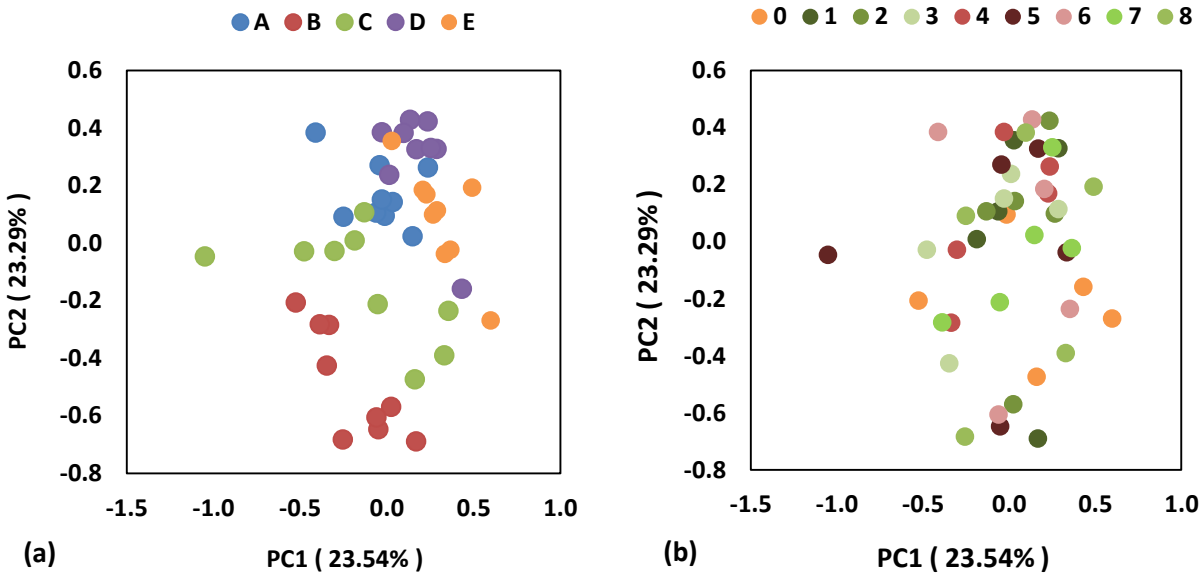


FIGURE 5.11 | Principal Component Analysis of Stool Microbiome Composition
Using the MG-RAST online platform, principal component analysis was completed, using taxonomy classifications from the RDP database, at a 97% alignment, with other parameters at default. Resulting plots show partial separation by (a) participant, particularly Participants A and C, but not by (b) sampling month. In (b), legend colour signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods.

TABLE 5.2 | Phylum and Genus Level Differences Between Stool Microbiome of Participants
A total of 38 genera, in four phyla, were shown to significantly differ between the five participants in the TAWT expedition. Significantly different genera are listed under their respective phyla (in bold) with groupings indicated by lowercase letters. Significant genera of unclassified taxonomy are not listed.

Taxonomic Level	A	B	C	D	E	*P* Value
Actinobacteria	**a**	**a**	**b**	**a**	**a**	**< 0.001**
Bifidobacterium	a	a	b	a	a	< 0.001
Collinsella	a	a,b	b	a	a,b	0.001
Enterorhabdus	a	a	b	a	a,b	0.005
Gordonibacter	a	a,b	b	a	a	0.011
Slackia	a	a	b	a	a	0.001
Bacteroidetes	**a**	**a**	**a**	**b**	**a**	**< 0.001**
Bacteroides	a	a	a	b	a	< 0.001
Barnesiella	a	a,b	a	a,c	a	0.026
Butyricimonas	a	a	a	a	b	0.039
Flavobacterium	a	a	a	a	b	< 0.001
Odoribacter	a	a	a	b	a	< 0.001
Parabacteroides	a	a	a	b	a	< 0.001
Prevotella	a	a	a,b	a	b	0.003
Firmicutes	**a**	**a**	**b**	**b**	**b**	**< 0.001**
Acetivibrio	a	a	a	a	b	< 0.001
Aeribacillus	a	b	a	a	a	< 0.001
Alkaliphilus	a	a	a	a	b	< 0.001
Anaerostipes	a	a	a	a	b	< 0.001
Blautia	a,b	b	a	a	a	0.005
Butyricicoccus	a	a	a,b	a	b	< 0.001
Clostridium	a	a,b	a,b	b	b	0.002
Desulfosporosinus	a	a	a,b	a	b	0.006
Desulfotomaculum	a	a,b	b	a,b	a,b	0.026
Dialister	a,b	a	a	b	c	< 0.001
Erysipelothrix	a	a	a	a	b	< 0.001
Ethanoligenens	a	a,b	b	a	a,b	0.005
Eubacterium	a	b,c	a,b	c	c	< 0.001
Faecalibacterium	a,b	a,b	a	b	b	0.002
Halobacillus	a	a	b	a	a	< 0.001
Heliobacillus	a	a	a	a	b	0.002
Lactobacillus	a	a	b	a	a	0.001
Peptoniphilus	a	a,b	b	a,b	a,b	0.013
Roseburia	a,b	a,b	b	a,b	a	0.042
Ruminococcus	a,d	b,c	b	a,d	d,c	< 0.001
Sarcina	a	a	a	a	b	0.002
Selenomonas	a	a,b	a,b	a	b	0.027
Tepidimicrobium	a	a	a	a	b	0.007
Thermoactinomyces	a	b	a	a	a,b	0.017
Tissierella	a	a	a	a	b	0.003
Proteobacteria	**a**	**a**	**a**	**a**	**b**	**< 0.001**
Rhodospirillum	a	a	a	b	a	0.015

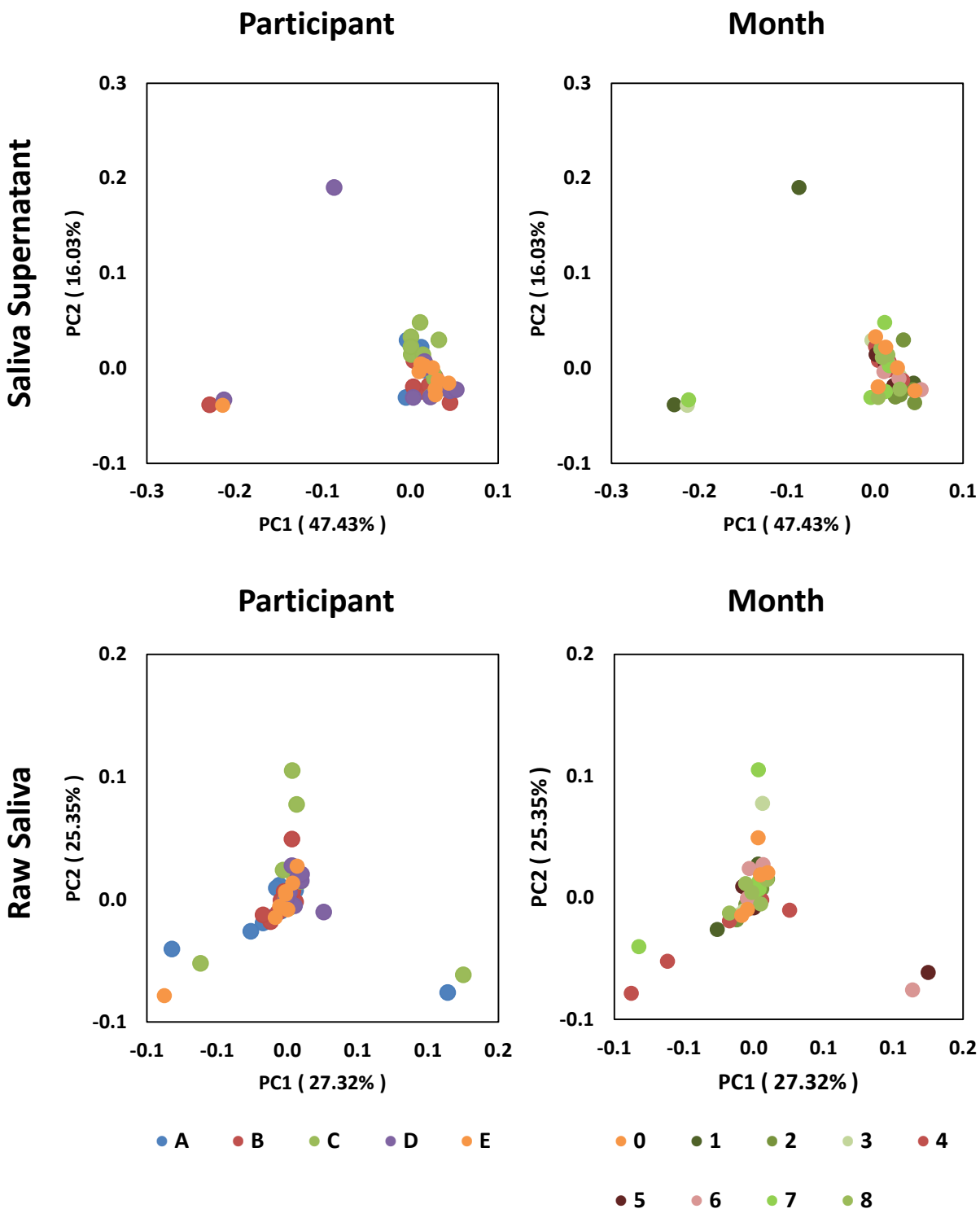


FIGURE 5.12 | Principal Component Analysis of Saliva Supernatant and Raw Saliva
Principal component analysis of negative mode LTQ-MS metabolite profiles was completed in PyChem, with samples classified by participant (left) and sampling month (right), for saliva supernatant (top) and raw saliva (bottom). Minimal variation was observed between participants, and no separation was observed between sampling months. In sampling month PCAs, legend colour signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods.

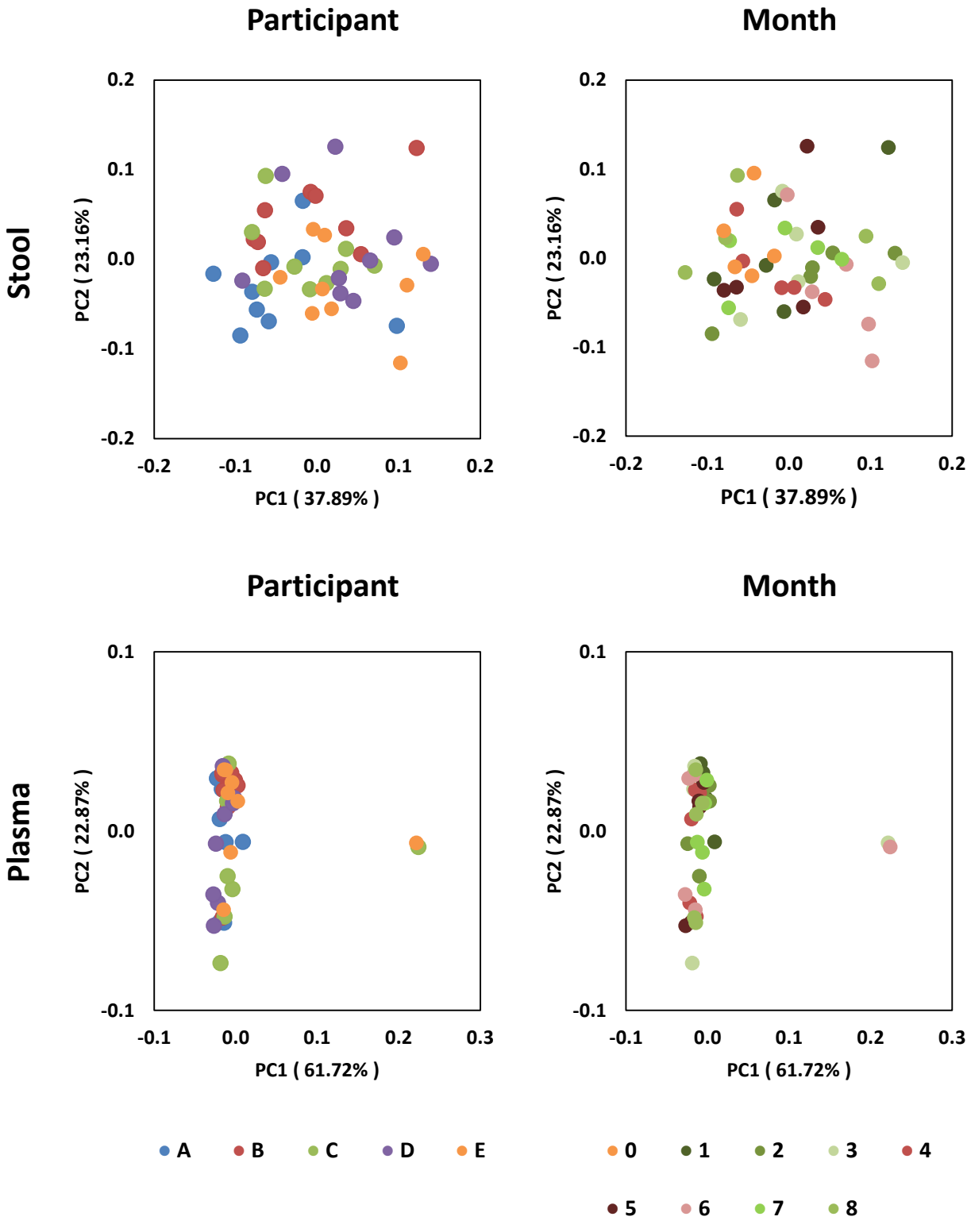


FIGURE 5.13 | Principal Component Analysis of Stool and Plasma
Principal component analysis of negative mode LTQ-MS metabolite profiles was completed in PyChem, with samples classified by participant (left) and sampling month (right), for stool (top) and plasma (bottom). For stool, some separation was evident between participants, but minimal variation was observed between sampling months. For plasma, some separation between participants was also observed, but not between sampling months. In sampling month PCAs, legend colour signifies activity of sampling period with orange showing baseline, green showing periods of traversing, and red showing stationary periods.

5.4 | Discussion

The human microbiome and metabolome are both important factors in maintaining host homeostasis. Dysbiosis associated with either could be important indicators of disease, or factors in their aetiology. Therefore, understanding how both are altered by extreme environmental and physiological conditions could be important in understanding how the human body would be affected by prolonged space travel, such as a mission to Mars. In this portion of work, the expedition members of the terrestrial Trans-Antarctic Winter Traverse, an analogous expedition to space travel, gave saliva, stool, and blood plasma samples over an eight month period, alongside an initial baseline sample. For all five participants, stool water content, biofluid pH, metabolome fingerprinting of biofluid, and bacterial load and diversity in saliva and stool was determined.

5.4.1 | Stool Water Content and Biofluid pH

Water is an essential nutrient in human health, and the most abundant compound in the human body. Water enters the human body through the consumption of fluid and solid food, and leaves via a number of pathways, including stool (Kleiner, 1999). Here, the water content of each monthly stool sample was determined by weight loss as a result of freeze drying. Overall, no significant differences were evident in regards to changes in stool water content between sampling months. Significant differences were however, seen between expedition members, suggesting individual differences in stool water content are maintained during physiological and environmental stress. The maintenance of hydration is important in health (Ritz and Berrut, 2005), and deviations from a person's normal stool water content could be an indicator of constipation, which has important health implications (Arnaud, 2003). Although stool water content appears to be unconnected with sampling month, in this portion of work, it is possible that water loss was altered elsewhere, such as increased urine volume.

As with stool water content, the pH of saliva supernatant, raw saliva, and blood plasma appears to maintain individual differences over the TAWT expedition, without being significantly affected by

sampling month. Significant differences were evident between the pH of participants' saliva supernatant and raw saliva, but not blood plasma. Due to its importance in metabolic function, particularly in enzyme function, the lack of individual differences in blood plasma pH is not unexpected. Changes in salivary pH however, are linked to the activation of the sympathetic nervous system, which controls the fight-or-flight response (Khalaila, Cohen and Zidan, 2014). Individual differences in salivary pH were maintained through the TAWT expedition and no differences between sampling months were evident, suggesting that stress-associated changes may not have been a substantial factor.

5.4.2 | Expedition Effects on the Salivary Microbiome

The salivary microbiome has not been widely studied in regards to the effect that prolonged environmental and physiological stress may have. In this portion of work, the bacterial load and diversity of the salivary microbiome was determined. In regards to bacterial load, no significant individual differences were evident between TAWT expedition members. Significant differences were however, present between sampling months, with baseline samples showing a lower bacterial load than at all other time points. The salivary bacterial load has previously been suggested as an *in vivo* marker of immunity (Jones *et al.*, 2014). At the time that baseline samples were donated, participants were not involved in the Antarctic traverse. By Month 1 however, the TAWT expedition had begun and participants were subjected to substantial environmental and physiological stress. The subsequent rise in salivary bacterial load could support the concept of it as an *in vivo* marker of immunity. However, because sample biofluids were unable to be stored at low ultralow temperatures, such as below -80°C, it was not plausible to accurately measure established markers of immunity, such as immunoglobulins or neutrophils (Vaught, 2006). If this had been possible, then establishing salivary bacterial load as an *in vivo* marker of immunity may have been possible. Salivary bacterial load has been linked, albeit not conclusively, to the onset of oral diseases (Dahan *et al.*, 2004). Bacterial load has however, been linked to a range of respiratory diseases, including COPD (Wilkinson *et al.*, 2003) and pneumonia (Muñoz-Almagro *et al.*, 2011). It may be that the increased bacterial load in saliva seen in this portion of work is

reflected in other portions of the respiratory tract. This could have potential implications for the health of participants in prolonged human space travel.

The bacterial diversity of the human salivary microbiome was measured through sequencing of the V3 to V4 region of the 16S rRNA gene. As a microbiome-wide measure of bacterial diversity, α-diversity values showed significant differences between participants, namely Participant A having an overall lower level of bacterial diversity, and sampling months, with baseline samples having lower bacterial diversity than all other time points. As with salivary bacterial load, the increase in bacterial diversity may be an indicator of altered immune function, particularly as the differences are seen in comparison to baseline samples and maintained through subsequent months of environmental and physiological stress. The role that bacterial diversity plays in the salivary microbiome is still to be established (Scannapieco, 2013). Some oral diseases, such as dental caries, have been linked to shifts in the taxonomic composition of the oral microbiome, rather than the emergence of distinct bacterial species (Yang *et al.*, 2012). Therefore, the increase in bacterial diversity seen here may be an indicator of dysbiosis in the microbiome, which could have important implications in terms of disease, and the host's health generally.

Bacterial load and diversity analysis appears to suggest that sampling month is an important factor in determining the bacterial composition of the salivary microbiome. Therefore, changes in specific taxonomic classifications, namely phylum and genus, were analysed. This revealed one phylum and six genera which showed significantly different abundance levels between at least two time points. The phylum Bacteroidetes showed a significant and substantial increase from baseline samples to all other sampling months, from approximately 1.5% to over 10% in total abundance. The only genus of Bacteroidetes, to show significant differences in sampling month, was *Prevotella*, which also displayed a significant increase from baseline samples to all other sampling time points, rising from approximately 1% to over 10% from baseline to Month 1 time points. Members of the *Prevotella* genus have been associated with a number of anaerobic infections, particularly in the oral cavity and respiratory tract

(Brook, 2007). Of the five other genera showing significant differences between sampling months, four are members of the Firmicutes phylum and one of the Actinobacteria phylum. The four members of the Firmicutes phylum, *Desulfotomaculum*, *Veillonella*, *Megasphaera*, and *Bacillus,* all show an increase in relative abundance from baseline samples to all other sampling months. The only genus to show a higher level of abundance in baseline samples was *Rothia*.

This portion of work appears to be unique in that it has profiled the bacterial load and diversity of the salivary microbiome in participants exposed to substantial environmental and physiological stress. To date, the focus in this field has been on establishing microbiome changes in the gastrointestinal tract which may impact human health on prolonged space travel. Due to its potential to impact or reflect the health of the host, this portion of work suggests that the salivary microbiome of participants in analogous situations to human travel should also be analysed.

5.4.3 | Maintenance of Individual Differences in Stool Microbiome

The human stool microbiome has been the focus of studies involving analogous situations to prolonged human space travel, including both human and animal-model studies. The stool microbiome of participants in the Mars-500 study, similar to the TAWT expedition, showed a substantial degree of stability in terms of function and maintenance of host health, albeit with some taxonomic changes (Mardanov *et al.*, 2013).

In this portion of work, the stool microbiome in both bacterial load and diversity showed a similar degree of stability, with individual differences between expedition members appearing to outweigh any alteration caused by sampling month. In regards to stool bacterial load, no significant differences were evident between participants or sampling month. Bacterial diversity in the form of α-diversity measures however, did show significant differences between participants, though no such differences were evident between sampling months. Decreased diversity of the intestinal microbiome, as represented by stool in this portion of work, has been associated with increased risk of developing intestinal conditions

including diarrheal disease (Bik and Relman, 2014), inflammatory bowel disease (Kinross, Darzi and Nicholson, 2011), and Crohn's disease (Clemente *et al.*, 2012). Therefore, the maintenance of bacterial load and diversity in the stool microbiome during the TAWT expedition suggests that participants were not at an increased risk of developing gastrointestinal disorders. With prolonged human space travel, such as a manned mission to Mars, the gastrointestinal microbiome would be subjected to substantial amounts of radiation. In mice, this has been shown to significantly alter the host's microbiome at both LD_{50} and LD_{30} levels (Karouia *et al.*, 2014). Although TAWT expedition members were not subjected to such radiation, the maintenance of stool bacterial diversity displayed does suggest its stability under substantial environmental and physiological conditions.

In addition to measures of bacterial diversity, namely α-diversity, macro level changes in the stool microbiome were modelled using principal component analysis. This supported previous results that individual differences are more substantial in determining the taxonomic composition of the stool microbiome than those resulting from sampling month. Principal component analysis by participant showed that Participants B and C were the most separate, though separations between Participants A, D, and E were also evident.

Due to the stool microbiome showing significant differences between participants, rather than sampling month, individual taxonomic differences were explored at the phylum and genus level of classification. This revealed a total of 38 genera, in four phyla, which showed significantly different abundances in at least one member of the TAWT expedition when compared to other members. The Bacteroidetes and Firmicutes phyla, which are usually the main constituents of the human gut microbiome (Lozupone *et al.*, 2012), displayed the greatest number of genera with significant differences between participants. The maintenance of these individual differences through the TAWT expedition suggest that they are an important component of stability and homeostasis in the gut microbiome, and that they are determined by other factors not including environmental and physiological stress.

Overall, the stool microbiome in the five TAWT expedition members appears to be highly stable, in terms of both bacterial load and diversity. Individual differences appear to be paramount in determining bacterial diversity, which persists even through sampling months where environmental and physiological stressors are extreme, with no overall difference in bacterial load between participants or sampling months. In a range of gastrointestinal diseases, and in general health, the human gut microbiome is clearly an important component. For some conditions, one of the key characteristics is a reduction in overall bacterial diversity in the gut (Kinross, Darzi and Nicholson, 2011). Here, bacterial diversity in the stool, shown through α-diversity, is significantly different between participants. The highest measure of α-diversity, in Participant C, was over twice that of the lowest measure, in Participant A. In terms of risk assessment for prolonged human space travel, the starting bacterial diversity in an individual's gut could potentially be used as a discriminating factor in selection of participants to reduce the likelihood of gastrointestinal issues during the mission.

5.4.4 | Stability of the Saliva, Stool and Plasma Metabolome

The metabolome of the four biofluids studied in this portion of work, using negative mode LTQ-MS, suggest that there is no overall difference between participants or sampling month. This suggests that there was no substantial, macro-level change in the metabolome of saliva, stool, or blood plasma samples from the five TAWT expedition members. Previous studies of the plasma metabolome in particular have suggested changes in certain metabolites associated with age, sex and race (Lawton *et al.*, 2008). Therefore, small level changes in the metabolome may have occurred as a result of the TAWT expedition, but because of the inherent similarities in participants, these were obscured.

The plasma metabolome has been shown to be altered by levels of physical fitness (Chorell *et al.*, 2012). Given that TAWT expedition members were of a similar level of physical fitness, to be able to meet the challenges posed by Antarctica, this could further obscure differences in the metabolome resulting from altered metabolism to maintain homeostasis. It may be that a more valid method of measuring changes in the human metabolome, either between participants or sampling month, would be to look at changes

in levels of individual metabolites. However, because of the sample size relative to the total number of metabolite peaks measured using negative mode LTQ-MS, the high rate of Type I statistical errors would make the subsequent findings unreliable (Broadhurst and Kell, 2006).

5.5 | Conclusions and Future Work

This portion of work aimed to take advantage of the undertaking of the TAWT expedition, which served as a unique opportunity to study human microbiome and metabolome changes in an analogous situation to prolonged human space travel, such as a manned mission to Mars. The TAWT expedition was ultimately unsuccessful in its intention to cross Antarctica during the winter months. Nevertheless, the extreme environmental and physiological stressors that expedition members experienced were still useful in modelling human microbiome and metabolome changes.

In regards to changes in the human microbiome, salivary bacterial load and diversity both increased markedly from baseline samples to the eight monthly samples collected during the TAWT expedition. Salivary bacterial load has previously been suggested as an *in vivo* marker of immune function, and it may be therefore, that immune function in TAWT expedition members was lowered, allowing for the observed increasing in bacterial load, and diversity. However, because the method of microbiome profiling employed here is unable to resolve taxonomic classification to the species level, it is difficult to conclude what effects these changes may have on host health and homeostasis.

The metabolome of saliva, stool and blood plasma showed no substantial differences between participants or sampling month. Stool water content, and the pH of raw saliva and saliva supernatant showed significant individual differences. However, analysis of changes in the human metabolome may have been hampered by the small sample size associated with the TAWT expedition.

Although this portion of work was limited by the relatively small sample size of participants, it has nevertheless created a framework on which subsequent studies into the real or analogous effect of prolonged human space travel can build upon. In regards to microbiome changes, subsequent work aiming to classify the saliva and stool microbiome in terms of its species-level composition and functional capacity could provide novel insights. Additionally, work on the analysis of individual metabolites, which would only be possible with larger sample sizes, is likely to provide useful insights.

CHAPTER 6 | General Discussion and Conclusions

The human microbiome and metabolome have been shown to be important components in disease, from aetiology to diagnosis, monitoring to treatment. In this research project, the aims were to address unanswered questions in both fields; and to combine the two to establish whether changes in one could explain changes in the other. The portions of work completed here are clearly divided into ones with a clinical application, Chapters 2 and 3, and those which answer more basic research questions, Chapters 4 and 5. Nevertheless, the tools and techniques which were developed during the Chapters with clinical applications could be easily applied to those answering basic research questions.

6.1 | Novel Insights in the Diseased Lung Microbiome and Metabolome

This research project was started with the intention of using metagenomic and metabolomic techniques to address unmet clinical needs that exist within the respiratory diseases lung cancer and COPD. The field of microbiomics is approximately ten years old, but little focus has been given to the lung microbiome; rather the gut microbiome has received the majority of the field's attention (Dickson, Erb-Downward and Huffnagle, 2013). Lung cancer in particular has only a few published studies on the lung microbiome in affected patients (Hosgood *et al.*, 2014; Laroumagne *et al.*, 2013; Koshiol *et al.*, 2012).

One of the primary aims of this research project was to address this using metagenomic sequencing, allowing the species-level resolution and functional capacity of the lung microbiome in lung cancer and COPD patients, represented by sputum samples, to be established. Additionally, sputum samples from lung cancer patients were to be analysed using metabolomic fingerprinting to identify non-invasive biomarkers for earlier detection of the disease.

In these regards, the project was successful. Albeit using a relatively small sample size, significant differences in both taxonomic composition and functional capability between lung cancer positive and lung cancer negative patients were detected. The use of this patient group is one of the key strengths of

the study. Lung cancer can present with a range of non-specific symptoms, which are shared with other non-cancerous lung conditions. Therefore, the ability to distinguish between patients with similar symptoms, but dissimilar causes, could have significant clinical implications. This was shown to be possible through analysis of specific features of the lung microbiome and metabolome. The apparent ability of the lung cancer microbiome to mirror disease staging, through levels of *G. adiacens* relative to six other bacterial species, using a non-invasive sampling method could aide in the monitoring of the disease without additional invasive procedures. However, the small sample size of patients, all of whom had NSCLC types, used in this pilot study potentially limits its future applicability to SCLC types. Further metagenomic sequencing with a larger patient cohort would be required to confirm whether these microbiome changes hold true. Nevertheless, in a clinical application a technique, such as quantitative PCR, which requires less resources could be utilised as a diagnostic tool. If sputum were to continue as the sampling biofluid, rather than the more invasive BAL or biopsy method, this could act as a non-invasive diagnostic or screening method; bringing the prospect of a targeted, at-risk population screen closer.

Using a larger sample cohort, metabolomic fingerprinting, whereby the structures of individual metabolites are not determined, identified a range of metabolites which are able to differentiate between patients positive and negative for lung cancer, but presenting with suspected lung cancer symptoms. As with microbiome-derived biomarkers, metabolomic fingerprinting allows for a potentially high-throughput, non-invasive method of diagnosis and screening. Nevertheless, this portion of work used a relatively small sample size. To support the use of mass-spectrometry metabolomics, and the use of sputum as a diagnostic medium, applicability of these findings to a larger cohort would need to be determined. Furthermore, the cohort used in this study contained a mix of lung cancer types, NSCLC and SCLC, and sub-types thereof, of varying disease stages. By using a larger cohort in further studies, the ability of mass-spectrometry metabolomic fingerprinting to differentiate between disease type and stage could be established. Because the biomarkers discovered in this portion of work are derived from

sputum, it is possible that the two approaches can be combined to increase the sensitivity and specificity rates of using sputum as a non-invasive medium for diagnosis and screening.

Chronic obstructive pulmonary disease is one of the few respiratory diseases that has received attention in regards to the lung microbiome in patients. However, these studies have used amplicon sequencing of the 16S rRNA gene to characterise the taxonomic composition of the microbiome (Erb-Downward *et al.*, 2011; Pragman *et al.*, 2012; Sze *et al.*, 2012). Although amplicon sequencing can give an overview of microbiome composition, it is limited in its ability to resolve bacteria to the species level of taxonomy, nor is it able to identify the functional capability of the microbiome. In this research project, the lung microbiome of patients with and without COPD was sequenced using metagenomics. This revealed both taxonomic and functional characteristics which differentiated the COPD microbiome, which could explain some of the clinical features of the disease including the rapid increase in bacterial load during exacerbations. When individual features of the lung microbiome in COPD were correlated with COPD severity, sialic acid metabolism was shown to have a positive correlation with FEV_1 % of predicted. Sialic acid is an important component of the inflammatory response, which is characteristic of COPD. This suggests that the lung microbiome in patients with COPD has a close interaction with the host's sialic acid, which could be an important mechanism in the pathogenesis of the disease. It could also provide a novel mechanism for therapeutic interventions and disease monitoring. However, as with the lung cancer microbiome and metabolome work, this study used a relatively small sample size, and further metagenomic sequencing of the lung microbiome in more COPD patients is required. Additionally, an important component of COPD is exacerbation of the condition, which is the common cause of death and worsening morbidity. This was not measured during this study. Longitudinal sampling of COPD patients would however, be able to establish whether the lung microbiome, as suggested here, was able to influence the progression of the disease. For example, the lung microbiome's capacity for rapid bacterial cell division and increased bacterial load could be used as a method for predicting a patient's risk of exacerbation.

6.2 | Temporal Variability of the Salivary Microbiome and Metabolome

One of the key questions in microbiome and metabolome research is whether they are stable over a prolonged period of time. This is of paramount importance where features of the microbiome and metabolome are used as biomarkers, for the screening, diagnosis, or monitoring of disease. In this portion of work, the salivary microbiome and metabolome, chosen because of the ease of sampling saliva, was monitored over a one year period, with samples taken every two months. Although other microbiome studies have studied the temporal variability of the human microbiome (Costello *et al.*, 2009), this has been over a relatively short period, frequently less than one month, or over long periods of time with large temporal gaps in sampling time points. However, if microbiome and metabolome biomarkers are to be used clinically, their variability over a yearly period needs to be determined. To this end, the portion of work outlined in Chapter 4 of this thesis has suggested that the salivary microbiome and metabolome is stable over a one year period, although bacterial load was significantly higher in the February sampling period than at all others.

This portion of work appears to be the first in the literature to longitudinally track the salivary microbiome and metabolome over a one year period, and to suggest that it is stable, except from the one variation in bacterial load seen in February. Due to limited resources being available for sequencing of the microbiome, metagenomic sequencing of all samples was not possible and amplicon sequencing of the 16S rRNA gene was chosen. This still gives a novel insight into the salivary microbiome, and indeed is still the most common method of microbiome sequencing in the literature (Caporaso *et al.*, 2012). Nevertheless, amplicon sequencing is unable to resolve to the species-level of taxonomy and it may be that temporal differences in the salivary microbiome are evident. Additionally, a large number of sequences from the salivary microbiome were of unclassified taxonomy. This is likely an issue with the difficulty in culturing bacteria from the human microbiome, and thus whole genome sequencing of isolated species has not been possible, leaving gaps in sequence databases. Undoubtedly, this will be solved by advances in culturing methods of bacteria (Vartoukian, Palmer and Wade, 2010), and by

improvements in our bioinformatic ability to assemble the genomes of mixed cultures into individual bacterial species, possibly through the use of single-cell genomics (Chitsaz *et al.*, 2011).

6.3 | White Mars – Stressing the Microbiome and Metabolome

Stress has a number of effects on the human body, and can be detrimental to health if homeostasis is unable to be maintained. An emerging area of research in the fields of microbiomics and metabolomics is how they are affected by stress, and if they are able to modulate the body's response to it. One such application of this emerging field is to prolonged human space travel, such as a manned mission to Mars. As an analogy to such a mission, the microbiome and metabolome of members of the TAWT expedition was monitored. The salivary microbiome showed significant changes between sampling months, whilst the stool microbiome showed the maintenance of individual differences throughout the expedition. Changes in the metabolome of saliva, stool, and blood plasma were not evident.

The seasonal variability of the salivary microbiome, as described in Chapter 4, suggests that the difference between sampling months during the TAWT expedition are as a consequence of the extreme physiological and environmental stress participants endured. To date, the focus of studies on how the human microbiome is affected by environmental and physiological stress has been the gastrointestinal tract. In this portion of work, spatial differences within the human microbiome have been shown suggesting the salivary microbiome changes as a result of environmental and physiological stress. Because the method of microbiome profiling employed, it is not possible to identify the species level changes within the microbiome; which may be able to reveal further novel information regarding the human microbiome's response to stress.

The TAWT expedition was a unique opportunity to monitor the effects of extreme physiological and environmental stress on the human microbiome and metabolome. Nevertheless, there were unavoidable weaknesses in its design; primarily the small, and limited to men, sample size. Additionally, the lack of resolution to the species-level of taxonomy, and the classification of the functional capability

of the microbiome may mean that changes in the stool microbiome, which appeared overall stable, were not possible to establish. Indeed, the functional capability of the stool may be a key component in the development of gastrointestinal disease (Qin *et al.*, 2010).

6.4 | Linking Microbiomics and Metabolomics

One of the aims of this study was to combine profiling of the human microbiome and metabolome to reveal novel insights into both, and determine the extent to which changes in one could be explained by changes in the other. Overall, this has not been accomplished in this portion of work.

In regards to the lung cancer and COPD microbiome and metabolome work, insufficient sample was available to allow for profiling of both. This meant that generating microbiome and metabolome profiles of sputum from the same patient was not possible, and therefore, any link between the two could not be established.

When the generation of microbiome and metabolome profiles from the same sample, such as in Chapters 4 and 5, was possible, changes in the microbiome did not appear to be explained by changes in the metabolome. Whilst charting the seasonal variability of the salivary microbiome for example, only seven genera were correlated with at least one metabolite. Although significant, these correlations were likely artefacts of the genera's low abundances. Within the White Mars portion of work, the sample size was determined to be too small to allow for statistically valid analysis between the microbiome and metabolome profiles generated.

Now that the field of microbiomics is moving towards a stronger focus on characterising the functional capacity of microbiomes, it is likely that metabolomics as a tool will be widely used. One of the key roles that metabolomics will play in understanding the human microbiome will be in analysing its metabolic products. The gut microbiome in particular has been shown to be enriched with genes that are adapted to the specific sources of energy that will be found there. However, the production of specific

metabolites by bacteria or the modulation of the host's metabolites may be an important component of diseases, such as Type II diabetes, which would not be realised through sequencing of the microbiome alone (Turnbaugh and Gordon, 2008).

6.5 | Separating Correlation and Causation

Changes in the taxonomic composition and functional capabilities of the human microbiome have been linked to a number of diseases (Cho and Blaser, 2012). In this research project, lung cancer has been linked to both and COPD to the microbiome. In COPD for example, the role that sialic acid metabolism may play is somewhat ambiguous, but could be fundamental in disease progression. However, because of the observational nature of both the lung cancer and COPD microbiome studies, only correlation and not causation can be established.

The issue of separating correlation and causation is not unique to this research project. In fact, it is one of the main issues facing the field of microbiomics. Although some diseases have been clearly linked to a microbial cause, such as *Helicobacter pylori* infection and gastric cancer (Uemura *et al.*, 2001), clearly linking disease-related changes in the microbiome with causing the disease itself is difficult without intervention studies. One approach which may prove effective is the use of model organisms. Although not all human diseases can be studied in a model organism, such as *Mus musculus*, many can. Additionally, the genetics of the model organism in relation to specific diseases is well understood, and the increasing commercial availability of germ-free *M. musculus* strains allows for intervention studies to be more widely completed. Nevertheless, intervention studies using model organisms may suffer from issues with translation to clinical intervention, but may still allow for a better understanding of disease dynamics (Cho and Blaser, 2012).

One of the key principals in microbiology is that of Koch's postulates, the four criteria which have to be met in order for the causative agent of a particular disease to be identified, and this will have to be applied to the field of microbiomics to fully separate cause and effect (de Vos and de Vos, 2012).

However, it may be that Koch's postulates do not hold true in all diseases linked to the microbiome. Because of its foundation in infectious disease, Koch's postulates focus on the identification of a single causative agent, such as a bacterium. It may be however, that certain diseases are caused by changes in more than one microbial species with a synergistic relationship.

6.6 | The Effect of Sample Choice, Storage, and Extraction

In studies of both the human microbiome and metabolome, the choice of sample is important. For example, it has been shown that sputum and BAL samples represent different spatial regions of the lung (Cabrera-Rubio *et al.*, 2012b), and therefore comparisons drawn between studies employing different sampling biofluids are likely to be limited.

Once the biofluid for sampling has been chosen, it may be that the methods for sample storage can impact upon the reliability and reproducibility of results. Analysis using NMR spectroscopy of the stability of human urine under room temperature (22°C), refrigeration (4°C), and deep-freeze (-80°C) has shown that approximately 50 metabolites change in concentration in room temperature and refrigerated samples, whilst those in samples in deep-freeze appear stable. This is may be as a result of bacterial contamination of the urine (Saude and Sykes, 2007).

Unlike the urine metabolome which shows different temporal changes depending on storage temperature, the microbiome of sputum from cystic fibrosis patients appears to be stable when a sample is stored at 4°C, -20°C or -80°C, but variation is introduced when the sample is stored at room temperature (approximately 25°C) (Zhao *et al.*, 2011).

For both microbiome and metabolome studies, sample storage at below 0°C appears to preserve the integrity of the sample. As both fields move towards increased standardisation of methods, sample storage conditions are likely to be fundamental in ensuring reproducibility and comparisons between studies are possible. However, the effect of long-term sample storage has not been firmly established.

Some biological assays, such as those measuring levels of thyroid hormones, sex hormones, and cytokines, have been successfully carried out on serum samples which have been stored for between 30 and 40 years (Jones and Golding, 2009). However, many longitudinal studies provide samples for analysis using technologies and techniques which were not available at its initiation. This may result in all of the samples from one longitudinal study being analysed in one batch, and therefore, any effect caused by sample storage duration would be difficult to establish.

In the body of work detailed here, a number of potential biomarkers derived from both the lung microbiome and metabolome have been identified. Although it appears that in samples stored at low temperatures, such as -80°C, the metabolome and microbiome profile is stable, it may be that variation is introduced within a particular microbial taxon or metabolite. Therefore, for any potential biomarker to be introduced into a clinical setting, it may be that the stability of individual biomarkers during storage needs to be established to ensure the reliability of a biomarker assay.

In studies of both the human microbiome and metabolome, the choice of extraction method for either DNA or metabolites respectively is important. For DNA extractions, a range of commercial kits and established laboratory techniques employing user-made buffers are available (Willner *et al.*, 2012a). Commercially available DNA extractions kits have been shown to vary in the estimated recovery of total genomic DNA from as little as 1% to up to approximately 40%, though reproducible profiles of bacterial communities may be possible between kits that employ bead-beating and lysis extraction methods (Vishnivetskaya *et al.*, 2014). Furthermore, analysis of bacterial community structure has been shown to be influenced by biological variation over technical variation in both human gut (Wagner Mackenzie, Waite and Taylor, 2015) and respiratory tract (Willner *et al.*, 2012a) samples. However, for samples of microbial communities with low biomass, it may be that DNA contamination present within commercially available kits introduces a more significant degree of technical variation, which may hinder the reproducibility of a study, and its comparison to others (Salter *et al.*, 2014)

Due to differences in sensitivity and detection range the choice of analysis method in metabolomic studies, such as between NMR or MS based approaches or different methods within each of these approaches, is likely to affect the view of the metabolome studied (Fiehn *et al.*, 2007). However, before samples can be analysed using either NMR or MS based approaches, an extraction must be completed; and the extraction method employed may also affect the view of the metabolome studied (Vuckovic, 2012). Extractions methodologies are usually based on solvents, such as ethanol, methanol, or chloroform, physical disruption, such as sonication or bead beating, or temperature, such as multiple freeze-thaw cycles. These various methods have been shown to not only extract differing numbers of metabolites, but also differing levels of the same metabolite from the same sample (Duportet *et al.*, 2011). This could have a significant effect on the biological interpretation of metabolomic based studies.

In this body of work, a number of biomarkers derived from the lung sputum metabolome have been suggested as having the potential to clinically discriminate patients with and without lung cancer. However, it is possible that unless the same method of chemical extraction is used, these metabolites may not be extracted with the same level of efficiency, and therefore, their discriminatory power may be limited. As the field of metabolomics moves towards standardisation, it may be that findings need to be reproducible on more than one platform of metabolomic profiling, such as GC-MS and LC-MS (Fiehn *et al.*, 2007). In addition to this, it may be that for findings to be valid the results have to be reproducible using more than one extraction methodology.

6.7 | Opportunities from Emerging Technologies

Advances in the capability of microbiome research can be directly correlated with advances in the capability of DNA sequencing, with the advent of high-throughput, next-generation sequencers, such as Roche's 454, being the initial catalyst. It is likely therefore, that future advances in the capability of microbiome research will be as a result of advances in sequencing technology. For example, nanopore DNA sequencing, such as Oxford Nanopore's pilot system, allows for substantially longer reads, with higher accuracy rates, than current sequencing systems (Bahassi and Stambrook, 2014). The use of this

system should allow for the sequencing of whole bacterial genomes, from a mixed microbial population, with sequence reads sufficient to allow for *in silico* genome assembly.

As a field, microbiomics is usually studied through the sequencing of isolated DNA. However, sequencing only reveals the potential metabolic functions of the microbiome, and does not give an understanding of how these may be altered under periods of stress, or disease. By combining metagenomic sequencing with other 'OMIC technologies, such as transcriptomics, proteomics, metabolomics, and lipidomics, for example, will allow for a measure of how the genetic and metabolic capacity of the microbiome translates *in vivo* (Serino, 2012).

The majority of research conducted into the human microbiome focuses on its bacterial component. This approach, however, fails to appreciate other microbial constituents, including eukaryotic microbes, protozoa, archaea, and viruses. The study design and analysis pipeline for these microbes is somewhat different than for bacteria. In regards to eukaryotic microbes, for example, no standardised marker gene exists and sequence databases are somewhat limited. In viruses particularly, no gene or genomic region is found across all viruses. Therefore, sequencing them alongside other microbiome constituents is not possible, and fractionation of virus-like particles is required. Many of the issues associated with the non-bacterial element of the microbiome will likely be solved through advances in sequencing, bioinformatic, and sequence repositories (Goodrich *et al.*, 2014).

6.8 | Developing Bioinformatic Capabilities and Techniques

Advances in our understanding of the human microbiome have largely been as a result of developments in sequencing capability. However, the rapidly increasing amount of data generated through sequencing studies needs to be matched by our ability to computationally and statistically analyse it (Chan and Ragan, 2013). This requirement is leading to the development of novel analysis pipelines and approaches. For example, in microbiome studies relying on sequencing of the 16S rRNA gene for taxonomic assignment, the issue of gene copy number may affect the reliability of quantitative

information. As a result, the tool CopyRighter has been developed to correct 16S rRNA amplicon microbiota profiles and quantitative abundances. This has been suggested to improve the comparability between amplicon and metagenomic data sets, and to improve estimates of bacterial diversity (Angly *et al.*, 2014).

The bacterial diversity of the microbiome's taxonomic composition has been one of the primary methods of analysis within the field. In this body of work, the measure of α-diversity, calculated by the MG-RAST pipeline using the Shannon diversity index, was frequently employed. This is a measure of the species richness within a given sample, namely, the higher the diversity value, the higher the number of bacterial species. For determining α-diversity, the Shannon and Simpson diversity indices are the two common methods employed. These two methods have been shown to have similar characteristics, but differ in the contribution of low-abundance taxa. Alternatives to these diversity indices include the Tail statistic, which shows greater sensitivity to low abundance taxa (Li *et al.*, 2012).

As with all other bioinformatic analysis techniques, choosing the appropriate tool is key in ensuring valid conclusions are drawn from a data set. In regards to diversity measurements, the Shannon and Simpson diversity indices may be more appropriate for some microbial communities, whilst the Tail statistic may be more useful when exploring communities with a high number of low abundance bacterial species.

The focus of bacterial diversity in microbiomics has been on taxonomic composition of the microbiome. However, as metagenomic sequencing becomes more prominent within the field, measures of community diversity may need to be developed to take in to account the greater detail and scope achieved using metagenomics. For example, these data sets allow for the diversity of not only taxonomy, but also of gene diversity, genotype diversity, phylogenetic diversity, evolutionary diversity, functional diversity, and structural diversity (Xu, 2006). It may be that for some diseases associated with dysbiosis in the human microbiome, gene diversity is a better measure than taxonomic diversity in terms of disease diagnosis or risk assessment.

With the cost of metagenomic sequencing declining, it is becoming a more accessible tool for analysis of the human microbiome. As a result, very large datasets are being created for each study, which increases the computational burden of sequence analysis. In addition, because metagenome sequencing usually take a shotgun approach, host sequence contamination is a more substantial issue than in amplicon sequencing. As a result, tools such as BMTagger have been developed to remove human reads from metagenomic datasets, which reduces the size, and thus computational requirements, of sequence files, and protects the identify of study participants (Gevers *et al.*, 2012).

For analysis of metagenomic sequence data, the assembly process is arguably one of the key steps. The Human Microbiome Project initially compared six assembly strategies and failed to identify one method with superior results (Gevers *et al.*, 2012). However, as the range of sequencing platforms has increased, each with its own advantages and drawbacks, assembly platforms have been developed to meet the specifications of each and the scope of the sequencing study. For example, the assembler Velvet has been developed for genome assemblies using Illumina, SOLiD, 454, and Sanger platforms, whilst ALLPATHS-LG is suited to Illumina and Pacific Biosciences platforms only. For metagenome assemblers, MetaVelvet, based on the Velvet genome assembler, has been developed for most sequencing platforms, whilst Genovo is specific to 454 pyrosequencing, and Meta-IDBA to Illumina only (The Human Microbiome Consortium, 2012a).

In comparison to amplicon sequencing datasets, metagenomic sequencing projects are able to reveal the metabolic and functional capacity of the human microbiome. However, this requires differing analysis pipelines to amplicon sequences, though methods in metagenomic sequence analysis are similar, in part, to those employed in whole genome studies. There are four main categories of methods employed in metagenomic sequence analysis to identify the functional assignment of genes. Firstly, homology-based approaches, which are used by the MG-RAST pipeline (Meyer *et al.*, 2008), are the most common. The other approaches include Motif- or pattern-based approaches, context-based annotation, and specific function approaches (Prakash and Taylor, 2012).

Metagenomic sequencing usually employs a shotgun-based approach, and because currently the widely used sequencing platforms, such as Illumina, are only able to generate short reads, the accuracy of gene prediction may be limited. With the increasing use of single cell genomics (Rinke *et al.*, 2013) and whole genome sequencing of bacterial isolates, this issue should be addressed. As it stands, metagenomic sequencing using current technology is likely to only be able to give a snapshot of the functional capacity of the microbiome, and is unlikely to be able to assign gene function to bacterial species (Prakash and Taylor, 2012).

Due to the limitations of the commonly employed sequencing platforms, which use short sequence reads, studies into the taxonomic composition of the human microbiome have been somewhat limited by the difficulty of resolving community structure to the species level. Indeed, this challenge is likely to be one of the key steps in moving beyond only the description of the microbiome and is becoming increasingly accessible through the use of metagenomic sequencing. The ability to resolve microbiome constituents to the species level of taxonomy will allow for a greater understanding of the evolutionary history of the microbiome. This will however, likely require a novel approach to the application of traditional evolutionary and ecological theory (Losos *et al.*, 2013).

During the analysis of metagenomic data sets, one of the key components is the prediction of gene function, and the assignment of the taxonomy of that gene. Currently, the main method of determining this is through similarity, such as using the BLAST algorithm. However, this approach can suffer from a number of drawbacks. For example, determining whether the sequence similarity between a gene of known function and a gene of unknown function is biologically significant is difficult. An alternative concept to assigning gene function based on similarity searches is that of using relatedness rather than similarity (Eisen and Wu, 2002).

The application of phylogenomics, which is the study of evolutionary relationships based on the comparative analysis of whole genome data, offers an alternative to traditional alignment-based

approaches to improve our understanding of the microbiome in terms of species relatedness and evolution. However, because the bacterial genomes which have, to date, undergone whole genome sequencing have been chosen because of their physiology, a highly biased phylogenetic distribution exists. Further whole genome sequencing of bacterial isolates, chosen because they cover a wide range of phylogenies, has recently been shown to allow for the discovery of novel protein families, improved gene function prediction, and the reconstruction of phylogenetic histories (Wu *et al.*, 2009).

The application of phylogenomics to metagenomic datasets requires significant computational resources, and it is plausible that with next-generation sequencing technology, traditional phylogenomics techniques may no longer be feasible. However, as with other bioinformatics techniques in microbiomics, those employed in phylogenomics are being adapted and developed for the characteristics of next-generation sequencing platforms. For example, phylogenomics has traditionally focussed on the use of multiple sequencing alignments for analysis. However, this technique is computationally intensive and may be infeasible for large datasets. As an alternative, alignment-free methods are being developed, such as those which employ *k*-mers, to reduce the requirement for intensive computation (Chan and Ragan, 2013).

To date, the field of microbiomics has developed concurrently with developments and advances in sequencing platforms. This is likely to continue to be the case for the foreseeable future, and additionally, the development of bioinformatic pipelines and analysis tools will be fundamental in exploiting advances in sequencing technologies. However, as the field expands to move beyond a descriptive account of the microbiome's constitution and functional capacity, there is likely to be a much wider variation in analysis techniques. This may continue to make standardisation within the field difficult.

CHAPTER 7 | Summary of Thesis Output

7.1 | Chapter 2 Output

A range of outputs were created as a result of the work detailed in Chapter 2. These are detailed here, including manuscript abstracts and graphical representations of posters for scientific conferences.

7.1.1 | Lung Cancer Microbiome Publication

Simon J. S. Cameron, Keir E. Lewis, Sharon A. Huws, Matthew J. Hegarty, Paul D. Lewis, Luis A. J. Mur, Justin A. Pachebat. (2015) Metagenomics of Lung Cancer Microbiome Suggests *Granulicatella adiacens* as a Biomarker for Status and Stage. *Under Review at Scientific Reports.*

ABSTRACT | Little is known about bacterial species-level composition and the functional capacity of the microbiome in patients with lung cancer (LC). Spontaneous sputum samples were collected from ten patients referred with possible LC, of which four were eventually diagnosed with LC (LC^+), and six had no LC after 1 year (LC^-). Genomic DNA was isolated and Nextera® metagenomic libraries constructed and sequenced on the HiSeq 2500 platform, with resulting sequences analysed using the MG-RAST metagenomic pipeline. Principal component analysis of taxonomic alignments showed some separation between LC^+ and LC^-, and was not influenced by smoking status. Of the seven bacterial species found in all samples, *Streptococcus viridans* was significantly higher in LC^+ samples. Seven further bacterial species were found only in controls, and 16 were found only in samples from LC^+. Additional taxonomic differences were identified in regards to significant fold changes between LC and controls cases, with five species having significantly higher abundances in LC^+. Functional differences, evident through significant fold changes, included polyamine metabolism and iron sidrephore receptors. Regression analyses correlated *Granulicatella adiacens* abundance with six other bacterial species in LC^+ samples. Further, bacterial species could also be related to LC stage. This study offers a novel insight into the functional capacity and species-level taxonomy of the sputum microbiome in patients with LC. Furthermore, it suggests *G. adiacens* as a novel and clinically useful biomarker for LC diagnosis and possible staging.

7.1.2 | Lung Cancer Metabolome Publication

Simon J. S. Cameron, Keir E. Lewis, Manfred Beckmann, Gordon G. Allison, Robin Ghosal, Paul D. Lewis, and Luis A. J. Mur. (2015) Metabolomic Fingerprinting of Clinical Sputum for Lung Cancer Biomarkers. *Manuscript in Preparation*.

ABSTRACT | Developing novel screening and diagnosis methodologies will allow for earlier diagnosis of lung cancer, increasing the effectiveness of clinical interventions. We tested the potential of metabolomic fingerprinting on sputum to differentiate between patients with and without lung cancer. We collected and processed the spontaneous sputum of 34 patients with suspected lung cancer, alongside 33 healthy controls. Of the 34 patients, 23 were subsequently diagnosed with lung cancer (16 NSCLC, six SCLC, and one radiological diagnosis) at various stages of disease progression. The 67 samples were analysed using linear quadrupole ion mass spectrometry (LTQ-MS) and gas-chromatography mass spectrometry (GC-MS). Principal component analysis clearly separated the clinically and non-clinically acquired samples using metabolites identified in negative LTQ-MS mode. Further, hierarchical cluster analysis, based on the top 25 metabolites identified through one-way ANOVAs in negative LTQ-MS mode, separated clinically and non-clinically acquired samples. Analysis based on area under the receiver operating characteristic curve (AUC) revealed differential metabolites for negative and positive lung cancer that had an AUC value of greater than 0.8. This preliminary analysis suggests sputum is a viable sample, and metabolomics has potential as a diagnostic and/or discriminator tool. Furthermore, it can identify specific key metabolites that could aid clinical intervention and targeted diagnosis of lung cancer within an 'at risk' population group. Metabolomics is a promising method for the identification of clinically useful biomarkers in the detection of lung cancer, but further work is required to establish biomarkers for stage and histology.

7.1.3 | Lung Cancer Diagnostic Patent Application

A patent application was made by Aberystwyth University to the UK Intellectual Property Office covering the work detailed in Chapter 2 on the use of microbiome biomarkers for the diagnosis and lung cancer stage and status. The patent application was made on 19th December 2014, with reference number P102979GB01.

PATENT APPLICATION ABSTRACT | The present invention relates to methods for the diagnosis of lung cancer (LC) status in a subject by determining the level of at least one specific 5 bacterial species relative to the level of at least one other bacterial species present in a sample which has been obtained from the subject (Fig. 3). The invention further provides a method of diagnosing LC stage in a subject by determining the level of at least one specific bacterial species in a sample which has been obtained from the subject, relative to the level of specific 10 bacterial species in a control. Also provided are kits, biomarkers and treatments for use in the methods described.

PRIFYSGOL ABERYSTWYTH UNIVERSITY

IBERS
Sefydliad y Gwyddorau Biolegol, Amgylcheddol a Gwledig
Institute of Biological, Environmental and Rural Sciences

Metagenomic Analysis of Sputum from COPD and Lung Cancer Patients Reveals Novel Insights in to the Structure and Function of the Upper Respiratory Microbiome

Simon J. S. Cameron[1*], Keir E. Lewis[2,3], Sharon A. Huws[1], Matthew J. Hegarty[1], Paul D. Lewis[3], Luis A. J. Mur[1†] and Justin A. Pachebat[1#].

[1]Institute of Biological, Environmental and Rural Sciences, Edward Llywd Building, Penglais Campus, Aberystwyth, SY23 3FG, UK. [2]Department of Respiratory Medicine, Prince Phillip Hospital, Llanelli, SA14 8LY, UK. [3]College of Medicine, Swansea University, Swansea, SA2 8PP, UK.

*sjc8@aber.ac.uk †lum@aber.ac.uk #jip@aber.ac.uk

Introduction

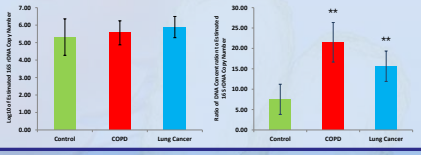
- Lung Cancer is the most prevalent cancer in the world, with 1.3 million deaths recorded each year. However, unlike most other cancers, it has seen very little change in its overall five year survival rate over the last 30 years[1].
- Worldwide, lung cancer and Chronic Obstructive Pulmonary Disease (COPD), a condition affecting the lungs ability to bring air in and out of the body, have a high level of morbidity and mortality, and share common risk factors in terms of tobacco smoking and a genetic predisposition[2].
- The microbiome of patients with COPD has been investigated[3], but as of yet, there has been no reported study of the microbiome in patients with lung cancer. In this study we aimed to address this lack of insight in to the lung microbiome of patients with lung cancer.

Study Methodology

- Spontaneous sputum samples were collected from a total of 20 clinical patients with either non-exacerbating-COPD (10) or lung cancer (10), and ten healthy age-matched individuals with no history of clinical lung disease.
- After DNA extraction using 100 µL of raw sputum, 50 ng of each sample was multiplexed using Illumina's Nextera DNA kit.
- Libraries then underwent paired-end sequencing using the Illumina HiSeq2500 rapid run platform with 2 x 151 bp reads, over two lanes. This generated approximately 10 million paired end reads per sample, after quality control using the MG-RAST metagenomic pipeline.
- Analysis of resulting sequencing statistics, post quality control, revealed no differences between the base pair counts ($p=0.418$) and sequence counts ($p=0.351$) for each of the three disease groups. The control group had significantly longer reads ($p=0.000$), by approximately ten bases, and the lung cancer group had a higher average GC content ($p=0.037$) by approximately 1.5 percentage points.

Analysis of Sputum Bacterial Load

- To gauge the overall bacterial load in the sputum of participants, quantitative PCR, targeting the 16S rDNA gene, using 2 µL of neat DNA was completed.
- No significant differences ($p=0.726$) were detectable between the three disease groups in terms of bacterial load. Additionally, no correlation was evident ($p=0.421$) between the concentration of extracted DNA and estimated 16S rDNA copy number.
- The ratio of DNA concentration to estimated 16S rDNA copy number is significantly increased in COPD and lung cancer patients ($p=0.004$), suggesting that the increased DNA has a non-bacterial source, such as increased bronchial cells in the case of COPD.

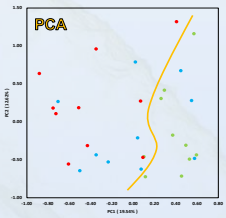


Conclusions

- This is the first piece of work to investigate the lung cancer microbiome, in addition to expanding our knowledge of COPD via analysis of the microbiome function.
- The structure of the microbiome in the three disease groups is similar; with no taxon being common amongst all samples in a group, and unique to that group.
- This work suggests that for COPD, the function of the microbiome is different to that of lung cancer and control samples, although the structure is the same.
- COPD functional differences appear to focus on an increased percentage of sequences related to cellular growth, indicating a potential for rapid colonisation, which may be an important factor in COPD exacerbations.

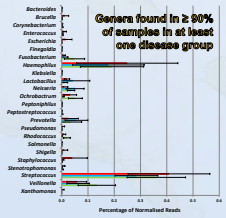
REFERENCES **(1)** Jemal A *et al.*, (2010) *CA: A Cancer Journal for Clinicians* 60: 277–300 **(2)** Young RP *et al.*, (2010) *The European Respiratory Journal* 36: 1375–1382. **(3)** Erb-Downward JR *et al.*, (2011) *PLoS ONE* 6(2): e16384.

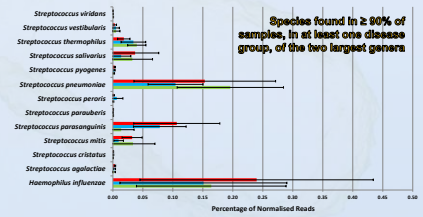
Taxonomic Analysis of Sputum Bacterial Communities



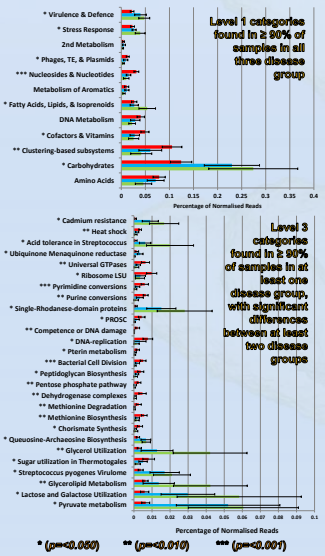
Figures Key: Control | Lung Cancer | COPD

- MG-RAST was used to align metagenomic sequences to the M5NR database, at a sequence identity of 97% and minimum alignment length of 15, to identify their taxonomic origins.
- As with other studies looking at taxonomic changes to the microbiome structure of the human respiratory system[3], our results support the observation that there is no genus, or species, that is common to all samples within a disease group, but unique to that disease group.
- At the genus level, only *Staphylococcus*, *Streptococcus*, *Ochrobactrum*, *Neisseria*, and *Pseudomonas* were found in all 30 samples. *Enterococcus*, *Lactobacillus*, *Veillonella*, *Fusobacterium*, *Brucella*, *Salmonella*, *Haemophilus*, and *Xanthomonas* were found in at least 90% of samples.
- Principal Component Analysis (PCA) suggests a partial separation between the control and the clinically acquired samples, particularly COPD, but this is by no means absolute.
- There were no taxons at which there were significant differences in the percentage of normalised reads between each disease group.



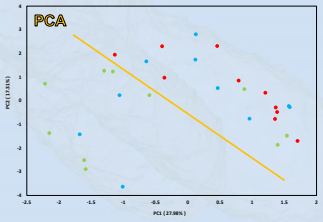


Functional Differences of Bacterial Communities



- MG-RAST was used to identify the functional role of metagenomic sequences using the Subsystems annotation source, with a sequence identity of 97% and minimum alignment length of 15.
- Although there appear to be no structural disease-caused changes to the microbiome, functional changes appear evident.
- COPD samples have increased percentages of normalised reads for Level 1 categories including those involved in nucleosides and nucleotides, DNA metabolism, cofactors and vitamins, and clustering-based subsystems.
- PCA suggests a partial separation between the control samples and the clinically acquired samples, but this is by no means absolute.
- COPD samples have reduced percentages of genes involved in heavy metal resistance, glycerol, lactose, and galactose utilisation, and pyruvate metabolism.
- In general, COPD samples show a significantly higher percentage of genes involved in growth, suggesting an increased capacity for microbial cell turnover and growth in patients with COPD.

Figures Key: Control | Lung Cancer | COPD



Metabolite fingerprinting of sputum targets biomarkers for the early identification of lung cancer patients

Simon J. S. Cameron[1*], Keir E. Lewis[2, 3], Manfred Beckmann[1], Gordon G. Allison[1], Robin Ghosal[2], Paul D. Lewis[3], and Luis A. J. Mur[1†].

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Ceredigion, SY23 3FG, UK. [2]Department of Respiratory Medicine, Prince Phillip Hospital, Llanelli, SA14 8LY, UK. [3]College of Medicine, Swansea University, Swansea, SA2 8PP, UK.

ABSTRACT

INTRODUCTION: Developing novel screening and diagnosis methodologies will allow earlier diagnosis of lung cancer (LC), increasing the effectiveness of clinical interventions. We tested the potential of the metabolomic approach known as metabolite fingerprinting of sputum to differentiate between patients with and without LC. Fingerprinting techniques use rapid high-throughput screening to identify variables which discriminate between samples without necessarily identifying them.

METHODS: Patients: We collected and processed the spontaneous sputum of 34 patients with suspected LC, alongside 33 healthy controls from staff at Swansea University, with no history of lung disease. Of the 34 patients, 23 were subsequently diagnosed with lung cancer (16 NSCLC, six SCLC, and one radiological diagnosis) at various stages of disease progression. **Procedure**: The 67 samples were profiled using linear quadrupole ion trap mass spectrometry (LTQ-MS) and gas-chromatography mass spectrometry (GC-MS).

RESULTS: Principal component analysis clearly separated the clinically and non-clinically acquired samples using metabolites in negative LTQ-MS mode. Further, hierarchical cluster analysis, based on the top 25 metabolites identified through one-way ANOVAs in negative LTQ-MS mode, separated clinically and non-clinically acquired samples. Analysis based on area under the receiver operating characteristic curve (AUC) revealed differential metabolites for negative and positive LC, with AUC values of greater than 0.8.

CONCLUSION: This preliminary analysis suggests that metabolomics could be used as a diagnostic and/or discriminator tool for LC diagnosis. Ultimately, it will identify key metabolites that could aid our understanding of the molecular pathogenesis of LC and possibly guide treatment targets.

BACKGROUND

LC is the most prevalent cancer in the world, and responsible for 1.3 million deaths each year[1]. The key to improving the five-year survival rate is the development of novel screening and diagnostic methodologies which allow for patients to be identified at an earlier stage in the disease, thereby increasing the clinical effectiveness of interventions.

Previous work by this research group has involved evaluating Fourier Transform infrared spectroscopy (FTIR), as a screening platform. FTIR was shown to differentiate between sputum from patients with and without LC. Furthermore, it supported the role that sputum could play as a biofluid for use in LC screening[2].

The aim of this study was to employ metabolomic fingerprinting to identify clinically relevant biomarkers, that could be used for the early detection of LC. This work is a foundation from which other studies, using larger numbers, could be based.

METHODS

SPONTANEOUS SPUTUM COLLECTION: 34 clinical patients (23 LC positive (LC+) (16 NSCLC, six SCLC, and one radiological diagnosis), and 11 negative (LC-)), in addition to 33 non-clinical controls (CON), (details in table, standard deviation in brackets, NC = not collected), gave sputum samples.

LTQ-MS PROFILING: A 50 % dilution of sputum underwent protein precipitation with acetone, then mixed with 70 % methanol before injection, in a randomised order, into a LTQ linear ion trap, in alternating positive and negative modes.

GC-MS PROFILING: For duplicate samples, protein precipitation was completed, as above. Derivitization was completed with methoxyamine and BSTFA, and run on an Agilent 6890N GC linked to a 5973N mass-spectrometer.

DATA ANALYSIS: Data was normalised, and analysed using PyChem[3], MetaboAnalyst 2.0[4], and ROCCET[5].

	CON	LC-	LC+
Number	33	11	23
Age	55.3 (14.6)	66.5 (14.3)	66.6 (8.1)
Gender			
Male	20	10	11
Female	13	1	12
Smoking Status			
Current	15	3	10
Ex	0	8	10
Never	18	0	3
Smoking Pack Years	NC	49.0 (34.9)	39.3 (18.9)
Infection Present			
Yes	NC	3	1
No	NC	8	22
CD Level (ppm)	NC	3.7 (1.3)	4.2 (2.8)

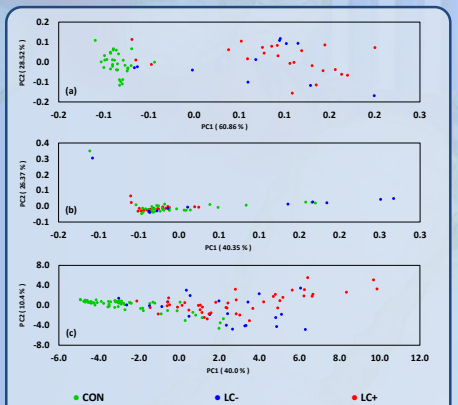
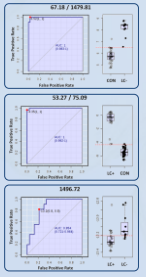


FIG 1: PRINCIPAL COMPONENT ANALYSIS (PCA) OF FINGERPRINTING METABOLITES
PCA plots were created for (a) negative and (b) positive LTQ-MS metabolites, and (c) GC-MS metabolites. The negative LTQ-MS profile showed the best separation, with the non-clinical controls being clearly different to the clinically acquire d samples.



	Metabolite	AUC Value	t-Test	Fold Change
CON Vs LC-	67.18/1479.81	1.00	4.47×10^{-15}	-2.38
	67.18/1560.81	1.00	3.10×10^{-15}	-2.26
	69.09/1479.81	0.99	6.09×10^{-15}	-2.25
	69.09/1560.81	0.99	3.64×10^{-15}	-2.14
	53.27/1209.45	0.99	1.79×10^{-14}	-2.28
LC+ Vs CON	53.27/75.09	1.00	8.74×10^{-24}	4.71
	69.09/75.09	1.00	7.42×10^{-24}	4.69
	189.09	1.00	6.12×10^{-20}	2.57
	187.09	0.99	1.42×10^{-18}	3.00
	190.00	0.99	3.03×10^{-20}	2.35
LC+ Vs LC-	1496.72	0.85	2.93×10^{-3}	-0.03
	957.36	0.83	4.57×10^{-3}	0.31
	1382.45	0.83	5.93×10^{-4}	0.07
	1396.27	0.82	1.28×10^{-2}	0.00
	1434.00	0.81	9.84×10^{-4}	0.14

FIG 4: RECEIVER OPERATING CHARACTERISTIC CURVE ANALYSIS FOR BIOMARKER IDENTIFICATION
Using the online facility ROCCET, univariate receiver operating characteristic curves (ROC) were created, and plotted to create area under the curve (AUC) figures for negative LTQ-MS metabolites. The top five AUC values for each comparison are given, showing a number of metabolites that could have a role as clinically useful biomarkers.

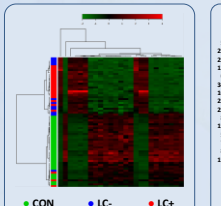


FIG 2: HIERARCHICAL CLUSTERING AND HEATMAP
Hierarchical cluster analysis, based on the top 25 negative LTQ-MS metabolites, identified by one-way ANOVAs, shows a similar degree of clustering as the PCA plot, with the clinical samples separating from non-clinical samples.

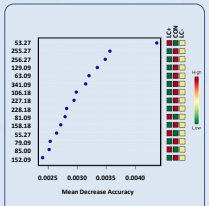


FIG 3: BIOMARKER DETECTION USING RANDOM FOREST
A Random forest plot was constructed, using MetaboAnalyst 2.0 for negative LTQ-MS metabolites, which revealed a number of metabolites, which may have potential in terms of diagnostic markers, particularly those that are either higher or lower in the LC+ group.

RESULTS & DISCUSSION

COMPARISON OF LTQ-MS AND GC-MS FINGERPRINTING APPROACHES: Metabolite fingerprints were modelled using principal component analysis (FIG 1), with metabolites in negative LTQ-MS mode (FIG 1a), showing the greatest degree of separation, though predominately between the clinically and non-clinically acquired samples. Metabolites identified in positive LTQ-MS (Fig 1b) and GC-MS (FIG1c) showed some degree of separation, though not as markedly as negative LTQ-MS metabolites. Similar separations were seen in hierarchical cluster analysis (FIG 2) with negative LTQ-MS metabolites.

TARGETING OF CLINICALLY RELEVANT METABOLOMIC BIOMARKERS: Because of the greater degree of separation shown using negative LTQ-MS metabolites, only these metabolites were used for the identification of potential biomarkers. A range of negative LTQ-MS metabolites (FIG 3 and 4) were suggested as potential biomarkers for LC status. Metabolites with very high AUC values, approximately 1.0, were differential between clinical and non-clinical samples. Additionally, metabolites with AUC values of greater than 0.8 were identified as differentials between LC+ and LC- groups.

DISCUSSION: Here we have reported on preliminary analysis suggesting that sputum is a viable diagnostic medium from which metabolite biomarkers could be used in the detection of LC. Specifically, we have identified a number of metabolites, with AUC values above 0.8, that are able to differentiate between positive and negative LC cases, in a clinical group of patients presenting with suspected LC. This suggests that this method may have a similar level of sensitivity and specificity as other methods, such as sputum cytology and bronchoscopy[6], without the need for an invasive procedure. This may make metabolomics an ideal, high-throughput tool for a preliminary negative/positive screen of 'at-risk' patients, which would identify them for further clinical diagnosis.

FUTURE PLANS: This preliminary study has suggested the potential application of metabolomics as a novel screening/diagnosis methodology for LC. Future work will concentrate on a larger number of clinical patients to establish whether metabolomics has the power to identify early-stage LC, and to differentiate histology.

REFERENCES

(1) WHO, 2013. Cancer Factsheet. *WHO Fact Sheets*, Number 2970. **(2)** Lewis, P.D., *et al.*, (2010). *BMC Cancer*. **(3)** Jarvis, R.M., *et al.*, (2006). *Bioinformatics* 22, 2565-6. **(4)** Xia, J., *et al.*, (2012). *Nucleic Acid Research*. 40, W127-33. **(5)** Xia, J., *et al.*, (2012). *Metabolomics* 9, 280-299. **(6)** Rivera *et al.*, (2013). *Chest Journal* 143, e142S-e165S.

* sjc8@aber.ac.uk or sjcameron@gmail.com † lum@aber.ac.uk

7.2 | Chapter 3 Output

A range of outputs were created as a result of the work detailed in Chapter 3. These are detailed here, including manuscript abstracts and graphical representations of posters for scientific conferences.

7.2.1 | COPD Metagenomics Publication

Simon J. S. Cameron, Keir E. Lewis, Sharon A. Huws, Matthew J. Hegarty, Paul D. Lewis, Luis A. J. Mur, Justin A. Pachebat. (2014) The Structure and Function of the Upper Respiratory Microbiome in Patients with Chronic Obstructive Pulmonary Disease Revealed Through Metagenomic Sequencing. *Under Review at Scientific Reports.*

ABSTRACT | The lung microbiome in COPD has been well-characterised, but little is known about the functional capacity of the microbiome that can be revealed by metagenomics. Genomic DNA was isolated from spontaneous sputa (sampling the upper respiratory tract) from ten control participants and eight patients with moderate to severe COPD. Barcoded Nextera® libraries were constructed from each sample and sequenced on the HiSeq 2500 platform. Resulting sequences were then analysed using the MG-RAST pipeline; an automated analysis platform for metagenomes. Principal component analyses showed partial separation for COPD based on for bacterial taxonomy and to a lesser extent based on function. Significant differences were observed in the abundance of bacterial species, particularly within the *Streptococcus* genus in COPD patients. The pathogenic species *Staphylococcus aureus*, *Stenotrophomonas maltophilia, Streptococcus agalactiae* and *Streptococcus pyogenes* were found in all COPD samples, but not all Control samples. Functional differences in the metagenomes from COPD patients were consistent with a greater bacterial growth capacity. Regression analyses correlated COPD severity with differences within the *Streptococcus* genus, specifically *Streptococcus pneumonia*. Metagenomic functional classifications indicated reduced bacterial sialic acid metabolism as COPD progressed. This pilot study has linked COPD severity to changes in abundance of *S. pneumonia* and reduced bacterial sialic acid metabolism which would allow host recognition to influence the

inflammatory response in the lung. Thus, the functional capacity of the COPD microbiome may be a factor in bacterial-associated progression and raises a number of avenues for further study including the novel COPD biomarkers.

PRIFYSGOL ABERYSTWYTH UNIVERSITY

IBERS
Sefydliad y Gwyddorau Biolegol, Amgylcheddol a Gwledig
Institute of Biological, Environmental and Rural Sciences

Metagenomics Reveals Insights into the Structure and Function of the Upper Respiratory Microbiome in Patients with Chronic Obstructive Pulmonary Disease

Simon J. S. Cameron[1]*, Keir E. Lewis[2,3], Sharon A. Huws[1], Matthew J. Hegarty[1], Paul D. Lewis[3], Luis A. J. Mur[1†] and Justin A. Pachebat[1]

[1] Institute of Biological, Environmental and Rural Sciences, Edward Llywd Building, Penglais Campus, Aberystwyth, SY23 3FG, UK. [2] Department of Respiratory Medicine, Prince Phillip Hospital, Llanelli, SA14 8LY, UK. [3] College of Medicine, Swansea University, Swansea, SA2 8PP, UK. ***Presenting Author: sjc8@aber.ac.uk †Communicating Author: lum@aber.ac.uk**

Introduction

- Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of death and morbidity worldwide, leading to an estimated 2.75 millions deaths worldwide in 2006[1].
- COPD is usually caused by smoking, and is an umbrella term for a multi-systemic inflammatory state, including diseases such as emphysema.
- Patients experience acute exacerbations, which may be triggered by infection or environmental pollutants. Around 75% are attributed to viral or bacterial infection, or a combination of both[2].
- The majority of microbiome work in COPD has employed sequencing of the 16S rRNA gene. However, minimal work has been undertaken to understand the function of the microbiome, and its species-level structure.

Materials and Methods

- Spontaneous sputum samples were collected from eight COPD patients (three moderate (GOLD II) and five severe (GOLD III)), and ten age-matched, controls, either present or ex smokers, with no history of lung disease.
- After DNA extraction using 100 µL of raw sputum, 50 ng of each sample was multiplexed using Illumina's Nextera DNA kit.
- Libraries then underwent paired-end sequencing using the Illumina HiSeq2500 rapid run platform with 2 x 151 bp reads, over two lanes. This generated approximately 10 million paired end reads per sample, after quality controls using the MG-RAST metagenomic pipeline.
- Further analysis was completed using the MG-RAST metagenomic pipeline, and MINITAB 14.

Core Species-Level Microbiome

- We found eight bacterial genera which were present in all 18 samples: *Haemophilus, Lactobacillus, Neisseria, Ochrobactrum, Pseudomonas, Staphylococcus, Streptococcus,* and *Veillonella.*
- At the species level (Figure 1), there are four species present in all 18 samples.
- Four additional species were found in all of the COPD samples but not all control samples – *Streptococcus aureus, Stenotrophomonas maltophilia, Streptococcus agalactiae*, and *Streptococcus pyogenes*, suggesting their potential to act as biomarkers of the disease.
- Significantly, all of these species are human pathogens, with *S. pyogenes* being the most pathogenic Streptococcal bacterium.

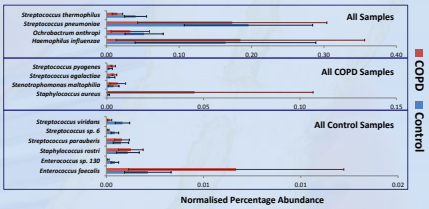


Figure 1 | Variations in species abundance in the 'Core Microbiome'

REFERENCES **(1)** Lopez AD *et* al., (2006) *Eur Respir J* 27: 397-412. **(2)** Han MK *et* al., (2012) *Thorax* 67: 456-463. **(3)** Bittar F and Rolain JM, (2010) *Clin Microbiol Infect* 16: 809-820. **(4)** Borges TJ *et al.*, (2012) *Front Immunol* 3: 95. **(5)** Cao H *et al.*, (2011) *Immunology* 132: 18-26.

Taxonomic Differences in the COPD Lung

- Some non-human eukaryotic sequences were identified, but no significant differences in species abundance were detected.
- Some species which exhibited statistically significant fold changes in abundance between COPD and controls were detected (Figure 2).
- The *Streptococcus* genus in particular appeared to exhibit many changes in abundance in COPD patient samples, with both increases and decreases compared to the Control group.
- These changes suggest that the lung microbiome in COPD is in flux, which may be characteristic of the altered environment in the lung.

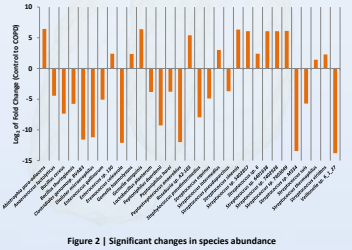


Figure 2 | Significant changes in species abundance

Functional Differences in the COPD Lung

- Significant increases in functional alignments appear to centre on those involved in bacterial growth, nucleosides and nucleotides, and amino acid, protein and RNA metabolism (Figure 3).
- Similar bacterial genomic flux also appears to be a feature of cystic fibrosis patients[3].
- Significant increase in alignments to the heat shock dnaK gene cluster, which in bacteria, is responsible for producing the heat shock protein Hsp70.
- Analogues of which have been shown to have significant anti-inflammatory responses[5], and could provide a mechanism for bacterial defence from the environment of the COPD lung.

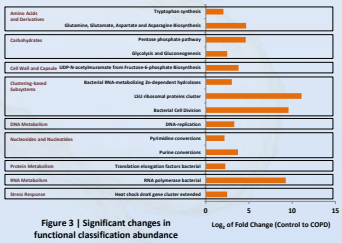


Figure 3 | Significant changes in functional classification abundance

Factors Associated with COPD Severity

- Significant positive correlations were found between COPD severity, based on FEV_1 % of predicted, with the *Streptococcus* genus, and more specifically *Streptococcus pneumonia*. Additional, functional positive correlations were observed with the Level 2 classification of di- and oligosaccharides, and at Level 3 with sialic acid metabolism.
- Additionally, these correlations were not observed in regards to smoking pack years, or age.
- *S. pneumonia* is commonly cultured from the sputum of patients during exacerbations, and is the main initial target of treatment with penicillin. Detecting changes in *S. pneumonia* load may allow for prediction of COPD progression and allow earlier interventions.
- Sialic acid are monosaccharides found commonly on the outside of vertebrate, and some microbial cells. Sialic acid interactions serve to maintain a baseline, non-activated state of innate immune cells, and limit inflammatory responses activation through PAMP/DAMP recognition[5].
- Here, sialic acid metabolism is correlated with decreasing FEV_1 scores which would indicate a shift towards a lesser capacity to avoid recognition and thus suppress inflammation – a key feature of COPD. Thus, a reduction in sialic acid metabolising bacteria could be an important pathological feature in COPD progression.

Table 1 | Regression analysis for COPD patients using FEV_1% of predicted, smoking pack years and age

Alignment Classification	Type (+ / -)	FEV_1 % of Predicted			Smoking Pack Years			Age		
		R^2	R^2 - adj	P	R^2	R^2 - adj	P	R^2	R^2 - adj	P
Streptococcus	+	51.8	43.7	0.044	23.1	10.3	0.228	20.6	7.3	0.259
Streptococcus pneumonia	+	63.6	57.6	0.018	8.4	0.0	0.487	40.1	30.1	0.092
Di- and oligosaccharides	+	50.8	42.6	0.047	1.8	0.0	0.752	29.7	18.0	0.162
Sialic Acid Metabolism	+	51.1	42.9	0.046	18.1	4.5	0.293	40.0	30.0	0.092

7.3 | Chapter 4 Output

The outputs from the work detailed in Chapter 4 of this thesis are detailed here.

7.3.1 | Human Salivary Microbiome Paper

Simon J. S. Cameron, Matthew J. Hegarty, Dan Smith, Luis A. J. Mur. (2015) The Human Salivary Microbiome Shows Temporal Stability in Bacterial Diversity but not Load. *Manuscript in Preparation*.

ABSTRACT | The human microbiome and metabolome are both important components of homeostasis, and in the monitoring of health and disease. However, the temporal stability of the human microbiome and metabolome has not been definitively established. Here, the saliva of 40 participants was collected every two months from October 2012 to October 2013, alongside lifestyle information. Samples were analysed through quantitative PCR to estimate bacterial load, measurement of pH, and metabolomic fingerprinting using negative mode LTQ-MS. A sub-group of ten participants with similar lifestyle information, underwent 16S rRNA (V3 to V4) amplicon sequencing. Estimated bacterial load was shown to be significantly (P < 0.001) higher in February 2013 than at all other sampling time points, with individuals' changes between time points displaying significant (P = 0.003) flux. Salivary pH levels were shown to be significantly (P < 0.001) higher in December 2012 than in October 2012 and February 2013, with significant (P < 0.001) individual variations seen across the sampling period. In regards to the stability of the taxonomic composition of the salivary microbiome, α-diversity values showed significant differences between participants (P < 0.001), but not between sampling periods (P = 0.801), and a small, but significant positive correlation with salivary pH (R^2 = 7.8%; P = 0.019). At the phylum level of classification, significant differences were evident between participants. The salivary metabolome also showed a strong degree of temporal stability. The salivary microbiome shows temporal stability in terms of bacterial taxonomic diversity, but not load, over a one year period, suggesting that comparing the human salivary microbiome and metabolome at different time points is valid.

7.4 | Chapter 5 Output

The outputs from the work detailed in Chapter 5 of this thesis are detailed here.

7.4.1 | White Mars Microbiome and Metabolome Paper

Simon J. S. Cameron, Arwyn Edwards, Matthew J. Hegarty, Mike A. Stroud, Alexander Kumar, Robert Lambert, Luis A. J. Mur. (2015) Humans and Their Hidden Companions Cross Antarctica: The Microbiome and Metabolome Effects of White Mars. *Manuscript in Preparation*.

ABSTRACT | The human microbiome and metabolome are important components of health and disease. The effects that prolonged human space travel could have on these are unclear, and could potentially impact the successful outcome of manned space travel, such as one to Mars. Here, we report on the microbiome and metabolome effects of the Trans-Antarctica Winter Traverse. Expedition members gave eight monthly stool, saliva, and blood plasma samples, and a preliminary baseline sample. Significant (P < 0.001) differences were seen between baseline salivary load and all other time points, but not between participants. Bacterial diversity was significantly (P = 0.002) lower in baseline samples than all other sampling months, though individual differences were also significant (P = 0.016). At the taxonomic level of classification, one phylum and six genera showed significantly different abundances between sampling months. The stool microbiome showed no difference in bacterial load between participants or sampling month, but bacterial diversity was significantly (P < 0.001) different between participants. Stool water content also showed significant (P < 0.001) differences between participants, as did pH of raw saliva (P < 0.001) and saliva supernatant (P = 0.001). In regards to metabolome changes, no difference was evident in any of the four biofluids profiled by either participant or sampling month. Overall, the human salivary microbiome shows significant and substantial changes in bacterial load and diversity as a result of the environmental and physiological stresses of the TAWT expedition. Thus, the human microbiome and metabolome is differentially affected based on site, which could have important implications for future, prolonged manned space travel to Mars.

7.5 | Non-Thesis Related Output

In addition to the work detailed within this thesis, additional work was also completed that has been published in peer-reviewed journals. These are outlined below, alongside a description of the contribution given to each.

7.5.1 | Jones *et al.*, (2014)

This research project investigated the physiological and immunological effects of supplementation with bovine colostrum in active males. Author contribution consisted of microbiome and metabolome work, including DNA extractions, 16S rRNA quantitative PCR, terminal restriction fragment length polymorphism analysis, metabolomic sample preparation, mass-spectrometry and analysis, subsequent data analysis and manuscript preparation.

Jones, A.W., **Cameron, S.J.S.**, Thatcher, R., Beecroft, M.S., Mur, L.A.J. and Davison, G. (2014). Effects of Bovine Colostrum Supplementation on Upper Respiratory Illness in Active Males. *Brain, Behaviour, and Immunity* 39, pp. 194–203.

ABSTRACT | Bovine colostrum (COL) has been advocated as a nutritional countermeasure to exercise-induced immune dysfunction and increased risk of upper respiratory illness (URI) in athletic populations, however, the mechanisms remain unclear. During winter months, under double-blind procedures, 53 males (mean training load±SD, 50.5±28.9 MET-hweek(-1)) were randomized to daily supplementation of 20g of COL (N=25) or an isoenergetic/isomacronutrient placebo (PLA) (N=28) for 12weeks. Venous blood was collected at baseline and at 12weeks and unstimulated saliva samples at 4 weeks intervals. There was a significantly lower proportion of URI days and number of URI episodes with COL compared to PLA over the 12weeks ($p<0.05$). There was no effect of COL on in vitro neutrophil oxidative burst, salivary secretory IgA or salivary antimicrobial peptides ($p>0.05$), which does not support previously suggested

mechanisms. In a subset of participants (COL=14, PLA=17), real-time quantitative PCR, targeting the 16S rRNA gene showed there was an increase in salivary bacterial load over the 12 weeks period with PLA ($p<0.05$) which was not as evident with COL. Discriminant function analysis of outputs received from serum metabolomics showed changes across time but not between groups. This is the first study to demonstrate that COL limits the increased salivary bacterial load in physically active males during the winter months which may provide a novel mechanism of immune-modulation with COL and a relevant marker of in vivo (innate) immunity and risk of URI.

7.5.2 | Edwards *et al.*, (2014)

This research project investigated the microbial and metabolomic component of cryoconite holes found on polar glacial surfaces and alpine regions. Author contribution consisted of sequence analysis generated from 454 pyrosequencing, and subsequent data analysis alongside relevant contributions to manuscript drafting.

Edwards, A., Mur, L.A.J., Girdwood, S.E., Anesio, A.M., Stibal, M., Rassner, S.M.E., Hell, K., Pachebat, J.A., Post, B., Bussell, J.S., **Cameron, S.J.S.**, Griffith, G.W., Hodson, A.J. and Sattler, B. (2014). Coupled Cryoconite Ecosystem Structure-Function Relationships are Revealed by Comparing Bacterial Communities in Alpine and Arctic Glaciers. *FEMS Microbiology Ecology* 89(2), pp. 222–237.

ABSTRACT | Cryoconite holes are known as foci of microbial diversity and activity on polar glacier surfaces, but are virtually unexplored microbial habitats in alpine regions. In addition, whether cryoconite community structure reflects ecosystem functionality is poorly understood. Terminal restriction fragment length polymorphism and Fourier transform infrared metabolite fingerprinting of cryoconite from glaciers in Austria, Greenland and Svalbard demonstrated cryoconite bacterial communities are closely correlated with cognate metabolite fingerprints. The influence of bacterial-

associated fatty acids and polysaccharides was inferred, underlining the importance of bacterial community structure in the properties of cryoconite. Thus, combined application of T-RFLP and FT-IR metabolite fingerprinting promises high throughput, and hence, rapid assessment of community structure–function relationships. Pyrosequencing revealed Proteobacteriawere particularly abundant, with Cyanobacteria likely acting as ecosystem engineers in both alpine and Arctic cryoconite communities. However, despite these generalities, significant differences in bacterial community structures, compositions and metabolomes are found between alpine and Arctic cryoconite habitats, reflecting the impact of local and regional conditions on the challenges of thriving in glacial ecosystems.

7.5.3 | Huws *et al.*, (2014)

This research project investigated the effect of flax and echium oil diet supplementation on the rumen lipidome and microbiome. Author contribution consisted of sequence analysis generated from 454 pyrosequencing and subsequent data analysis, alongside relevant contribution to manuscript preparation.

Huws, S.A., Kim, E.J., **Cameron, S.J.S.**, Girdwood, S.E., Davies, L., Tweed, J., Vallin, H. and Scollan, N.D. (2014). Characterization of the Rumen Lipidome and Microbiome of Steers Fed a Diet Supplemented with Flax and Echium Oil. *Microbial Biotechnology*. In Press.

ABSTRACT | Developing novel strategies for improving the fatty acid composition of ruminant products relies upon increasing our understanding of rumen bacterial lipid metabolism. This study investigated whether flax or echium oil supplementation of steer diets could alter the rumen fatty acids and change the microbiome. Six Hereford × Friesian steers were offered grass silage/sugar beet pulp only (GS), or GS supplemented either with flax oil (GSF) or echium oil (GSE) at 3% kg−1 silage dry matter in a 3 × 3 replicated Latin square design with 21-day periods with rumen samples taken on day 21 for the analyses

of the fatty acids and microbiome. Flax oil supplementation of steer diets increased the intake of polyunsaturated fatty acids, but a substantial degree of rumen biohydrogenation was seen. Likewise, echium oil supplementation of steer diets resulted in increased intake of 18:4n-3, but this was substantially biohydrogenated within the rumen. Microbiome pyrosequences showed that 50% of the bacterial genera were core to all diets (found at least once under each dietary intervention), with 19.10%, 5.460% and 12.02% being unique to the rumen microbiota of steers fed GS, GSF and GSE respectively. Higher 16S rDNA sequence abundance of the genera Butyrivibrio, Howardella, Oribacterium, Pseudobutyrivibrio and Roseburia was seen post flax feeding. Higher 16S rDNA abundance of the genus Succinovibrio and Roseburia was seen post echium feeding. The role of these bacteria in biohydrogenation now requires further study.

7.5.4 | Hadfield *et al.*, (2015)

This research project aimed to develop a method of obtaining whole genome sequences of *Cryptosporidium* isolates from clinical samples, without the requirement to grow the parasite in an animal model. Author contribution consisted of developing a 16S rRNA qPCR screen to detect levels of bacterial contamination, and efficacy of purification methods to remove bacterial contamination.

Hadfield S.J., Pachebat J.A., Swain M.T., Robinson G., **Cameron S.J.S.**, Alexander J.L., Hegarty M., Elwin K., Chalmers R.M. (2015). Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *Submitted to BMC Genomics.*

ABSTRACT | Whole genome sequencing (WGS) of *Cryptosporidium* spp. has previously relied on propagation of the parasite in animals to generate enough oocysts from which to extract DNA of sufficient quantity and purity for analysis. We have developed and validated a method for preparation of genomic *Cryptosporidium* DNA suitable for WGS directly from human stool samples and used it to

generate 10 high quality whole *Cryptosporidium* genome assemblies. Our method uses a combination of salt flotation, immunomagnetic separation (IMS), and surface sterilisation of oocysts prior to DNA extraction, with subsequent use of the transposon-based Nextera XT kit to generate libraries for sequencing on Illumina platforms. IMS was found to be superior to caesium chloride density centrifugation for purification of oocysts from small volume stool samples and for reducing levels of contaminant DNA. The IMS-based method was used initially to sequence whole genomes of *Cryptosporidium hominis gp60* subtype IbA10G2 and *Cryptosporidium parvum gp60* subtype IIaA19G1R2 from small amounts of stool left over from diagnostic testing of clinical cases of cryptosporidiosis. The *C. parvum* isolate was sequenced to a mean depth of 51.8x with reads covering 100% of the bases of the *C. parvum* Iowa II reference genome (Bioproject PRJNA 15586), while the *C. hominis* isolate was sequenced to a mean depth of 34.7x with reads covering 98% of the bases of the *C. hominis* TU502 v1 reference genome (Bioproject PRJNA 15585). The method was then applied to a further 17 stools, successfully generating another eight new whole genome sequences, of which two were *C. hominis* (*gp60* subtypes IbA10G2 and IaA14R3) and six *C. parvum* (*gp60* subtypes IIaA15G2R1 from three samples, and one each of IIaA17G1R1, IIaA18G2R1, and IIdA22G1), demonstrating the utility of this method to sequence *Cryptosporidium* genomes directly from clinical samples. This development is especially important for *C. hominis* for which symptomatic animal models, producing large numbers of oocysts, are lacking. This represents the first report of high quality whole genome sequencing of *Cryptosporidium* isolates prepared directly from human stool samples.

CHAPTER 8 | References

Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I. and Dewhirst, F.E., 2005. Defining the Normal Bacterial Flora of the Oral Cavity. *Journal of Clinical Microbiology*, 43(11), pp.5721–32.

Agusti, A. and MacNee, W., 2013. The COPD Control Panel: Towards Personalised Medicine in COPD. *Thorax*, 68(7), pp.687–90.

Ahn, J., Chen, C.Y. and Hayes, R.B., 2012. Oral Microbiome and Oral and Gastrointestinal Cancer Risk. *Cancer Causes and Control*, 23(3), pp.399–404.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), pp.403–10.

Andermann, A., Blancquaert, I., Beauchamp, S., Déry, V. and Dery, V., 2008. Revisiting Wilson and Jungner in the Genomic Age: A Review of Screening Criteria Over the Past 40 Years. *Bulletin of the World Health Organization*, 86(4), pp.317–319.

Angly, F.E., Dennis, P.G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P. and Tyson, G.W., 2014. CopyRighter: A Rapid Tool for Improving the Accuracy of Microbial Community Profiles Through Lineage-Specific Gene Copy Number Correction. *Microbiome*, 2, p.11.

Arnaud, M., 2003. Mild Dehydration: A Risk Factor of Constipation? *European Journal of Clinical Nutrition*, 57 Suppl 2, pp.S88–95.

Bach, P.B., Mirkin, J.N., Oliver, T.K., Azzoli, C.G., Berry, D.A., Brawley, O.W., Byers, T., Colditz, G.A., Gould, M.K., Jett, J.R., Sabichi, A.L., Smith-Bindman, R., Wood, D.E., Qaseem, A. and Detterbeck, F.C., 2012. Benefits and Harms of CT Screening for Lung Cancer: A Systematic Review. *JAMA*, 307(22), pp.2418–29.

Bahassi, E.M. and Stambrook, P.J., 2014. Next-Generation Sequencing Technologies: Breaking the Sound Barrier of Human Genetics. *Mutagenesis*, 29(5), pp.303–10.

Baldwin, D.R., Duffy, S.W., Wald, N.J., Page, R., Hansell, D.M. and Field, J.K., 2011. UK Lung Screen (UKLS) Nodule Management Protocol: Modelling of a Single Screen Randomised Controlled Trial of Low-Dose CT Screening for Lung Cancer. *Thorax*, 66(4), pp.308–313.

Barrett, P. and Bolborea, M., 2012. Molecular Pathways Involved in Seasonal Body Weight and Reproductive Responses Governed by Melatonin. *Journal of Pineal Research*, 52(4), pp.376–88.

Beaugerie, L., Seksik, P., Nion-Larmurier, I., Gendre, J.-P. and Cosnes, J., 2006. Predictors of Crohn's Disease. *Gastroenterology*, 130(3), pp.650–6.

Beyoğlu, D. and Idle, J.R., 2013. Metabolomics and its Potential in Drug Development. *Biochemical Pharmacology*, 85(1), pp.12–20.

Bik, E.M. and Relman, D.A., 2014. Unrest at Home: Diarrheal Disease and Microbiota Disturbance. *Genome Biology*, 15(6), p.120.

Bisgaard, H., Hermansen, M.N., Buchvald, F., Loland, L., Halkjaer, L.B., Bonnelykke, K., Brasholt, M., Heltberg, A., Vissing, N.H., Thorsen, S.V., Stage, M. and Pipper, C.B., 2007. Childhood Asthma After Bacterial Colonization of the Airway in Neonates. *New England Journal of Medicine*, 357(15), pp.1487–1495.

Bittar, F. and Rolain, J.M., 2010. Detection and Accurate Identification of New or Emerging Bacteria in Cystic Fibrosis Patients. *Clinical Microbiology and Infection*, 16(7), pp.809–20.

Bizzarro, M.J., Callan, D.A., Farrel, P.A., Dembry, L.M. and Gallagher, P.G., 2011. *Granulicatella adiacens* and Early-Onset Sepsis in Neonate. *Emerging Infectious Diseases*, 17(10), pp.1971–3.

Bogdanov, M., Matson, W.R., Wang, L., Matson, T., Saunders-Pullman, R., Bressman, S.S. and Flint Beal, M., 2008. Metabolomic Profiling to Develop Blood Biomarkers for Parkinson's Disease. *Brain*, 131(Pt 2), pp.389–96.

Bonne, N.J. and Wong, D.T., 2012. Salivary Biomarker Development Using Genomic, Proteomic and Metabolomic Approaches. *Genome Medicine*, 4(10), p.82.

Borges, T.J., Wieten, L., van Herwijnen, M.J.C., Broere, F., van der Zee, R., Bonorino, C. and van Eden, W., 2012. The Anti-Inflammatory Mechanisms of Hsp70. *Frontiers in Immunology*, 3, p.95.

Bradshaw, D.J., Homer, K.A., Marsh, P.D. and Beighton, D., 1994. Metabolic Cooperation in Oral Microbial Communities During Growth on Mucin. *Microbiology*, 140(12), pp.3407–3412.

Broadhurst, D.I. and Kell, D.B., 2006. Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. *Metabolomics*, 2(4), pp.171–196.

Brook, I., 2007. The Role of Anaerobic Bacteria in Upper Respiratory Tract and Other Head and Neck Infections. *Current Infectious Disease Reports*, 9(3), pp.208–217.

Byrne, M.M., Weissfeld, J. and Roberts, M.S., 2008. Anxiety, Fear of Cancer, and Perceived Risk of Cancer following Lung Cancer Screening. *Medical Decision Making*, 28(6), pp.917–925.

Cabrera-Rubio, R., Collado, M.C., Laitinen, K., Salminen, S., Isolauri, E. and Mira, A., 2012a. The Human Milk Microbiome Changes Over Lactation and is Shaped by Maternal Weight and Mode of Delivery. *The American Journal of Clinical Nutrition*, 96(3), pp.544–51.

Cabrera-Rubio, R., Garcia-Núñez, M., Setó, L., Antó, J.M., Moya, A., Monsó, E. and Mira, A., 2012b. Microbiome Diversity in the Bronchial Tracts of Patients with Chronic Obstructive Pulmonary Disease. *Journal of Clinical Microbiology*, 50(11), pp.3562–8.

Cao, H. and Crocker, P.R., 2011. Evolution of CD33-Related Siglecs: Regulating Host Immune Functions and Escaping Pathogen Exploitation? *Immunology*, 132(1), pp.18–26.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R., 2010. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nature Methods*, 7(5), pp.335–6.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G. and Knight, R., 2012. Ultra-High-Throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms. *The ISME Journal*, 6(8), pp.1621–4.

Carlin, A.F., Lewis, A.L., Varki, A. and Nizet, V., 2007. Group B Streptococcal Capsular Sialic Acids Interact with Siglecs (Immunoglobulin-Like Lectins) on Human Leukocytes. *Journal of Bacteriology*, 189(4), pp.1231–7.

Cassidy, A., Myles, J.P., van Tongeren, M., Page, R.D., Liloglou, T., Duffy, S.W. and Field, J.K., 2008. The LLP Risk Model: An Individual Risk Prediction Model for Lung Cancer. *British Journal of Cancer*, 98(2), pp.270–276.

Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A. and Holt, R.A., 2012. *Fusobacterium nucleatum* Infection is Prevalent in Human Colorectal Carcinoma. *Genome Research*, 22(2), pp.299–306.

Chan, C.X. and Ragan, M.A., 2013. Next-Generation Phylogenomics. *Biology Direct*, 8, p.3.

Chancellor, J.C., Scott, G.B.I. and Sutton, J.P., 2014. Space Radiation: The Number One Risk to Astronaut Health beyond Low Earth Orbit. *Life*, 4(3), pp.491–510.

Chang, Y.-C. and Nizet, V., 2014. The Interplay Between Siglecs and Sialylated Pathogens. *Glycobiology*.

Charlson, E.S., Bittinger, K., Haas, A.R., Fitzgerald, A.S., Frank, I., Yadav, A., Bushman, F.D. and Collman, R.G., 2011. Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. *American Journal of Respiratory and Critical Care Medicine*, 184(8), pp.957–63.

Cheema, A.K., Suman, S., Kaur, P., Singh, R., Fornace, A.J. and Datta, K., 2014. Long-Term Differential Changes in Mouse Intestinal Metabolomics After γ and Heavy Ion Radiation Exposure. *PLoS One*, 9(1), p.e87079.

Chen, W., Liu, F., Ling, Z., Tong, X. and Xiang, C., 2012. Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer. *PLoS One*, 7(6), p.e39743.

Chen, Y., Ma, Z., Li, A., Li, H., Wang, B., Zhong, J., Min, L. and Dai, L., 2014. Metabolomic Profiling of Human Serum in Lung Cancer Patients Using Liquid Chromatography/Hybrid Quadrupole Time-of-Flight Mass Spectrometry and Gas Chromatography/Mass Spectrometry. *Journal of Cancer Research and Clinical Oncology*.

Chitsaz, H., Yee-Greenbaum, J.L., Tesler, G., Lombardo, M.-J., Dupont, C.L., Badger, J.H., Novotny, M., Rusch, D.B., Fraser, L.J., Gormley, N.A., Schulz-Trieglaff, O., Smith, G.P., Evers, D.J., Pevzner, P.A. and Lasken, R.S., 2011. Efficient *de novo* Assembly of Single-Cell Bacterial Genomes from Short-Read Data Sets. *Nature Biotechnology*, 29(10), pp.915–21.

Cho, I. and Blaser, M.J., 2012. The Human Microbiome: At the Interface of Health and Disease. *Nature Reviews: Genetics*, 13(4), pp.260–70.

Chorell, E., Svensson, M.B., Moritz, T. and Antti, H., 2012. Physical Fitness Level is Reflected by Alterations in the Human Plasma Metabolome. *Molecular bioSystems*, 8(4), pp.1187–96.

Ciofu, O., Hansen, C.R. and Høiby, N., 2013. Respiratory Bacterial Infections in Cystic Fibrosis. *Current Opinion in Pulmonary Medicine*, 19(3), pp.251–8.

Claudino, W.M., Goncalves, P.H., di Leo, A., Philip, P.A. and Sarkar, F.H., 2012. Metabolomics in Cancer: a Bench-to-Bedside Intersection. *Critical Reviews in Oncology and Hematology*, 84(1), pp.1–7.

Clayton, T.A., Lindon, J.C., Cloarec, O., Antti, H., Charuel, C., Hanton, G., Provost, J.-P., Le Net, J.-L., Baker, D., Walley, R.J., Everett, J.R. and Nicholson, J.K., 2006. Pharmaco-Metabonomic Phenotyping and Personalized Drug Treatment. *Nature*, 440(7087), pp.1073–7.

Clemente, J.C., Ursell, L.K., Parfrey, L.W. and Knight, R., 2012. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell*, 148(6), pp.1258–70.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M. and Tiedje, J.M., 2009. The Ribosomal Database Project: Improved Alignments and New Tools for rRNA Analysis. *Nucleic Acids Research*, 37(Database issue), pp.D141–5.

Collins, S.M., Surette, M. and Bercik, P., 2012. The Interplay Between the Intestinal Microbiota and the Brain. *Nature Reviews: Microbiology*, 10(11), pp.735–42.

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L., 2009. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *International Journal of Epidemiology*, 38(5), pp.1175–1191.

Corona, G., Rizzolio, F., Giordano, A. and Toffoli, G., 2012. Pharmaco-Metabolomics: An Emerging 'omics' Tool for the Personalization of Anticancer Treatments and Identification of New Valuable Therapeutic Targets. *Journal of Cellular Physiology*, 227(7), pp.2827–31.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I. and Knight, R., 2009. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*, 326(5960), pp.1694–7.

Cryan, J.F. and Dinan, T.G., 2012. Mind-Altering Microorganisms: The Impact of the Gut Microbiota on Brain and Behaviour. *Nature Reviews: Neuroscience*, 13(10), pp.701–12.

Cryan, J.F. and O'Mahony, S.M., 2011. The Microbiome-Gut-Brain Axis: From Bowel to Behaviour. *Neurogastroenterology and Motility*, 23(3), pp.187–92.

D'Urso, V., Doneddu, V., Marchesi, I., Collodoro, A., Pirina, P., Giordano, A. and Bagella, L., 2013. Sputum Analysis: Non-Invasive Early Lung Cancer Detection. *Journal of Cellular Physiology*, 228(5), pp.945–51.

Dahan, M., Timmerman, M.F., Van Winkelhoff, A.J. and Van der Velden, U., 2004. The Effect of Periodontal Treatment on the Salivary Bacterial Load and Early Plaque Formation. *Journal of Clinical Periodontology*, 31(11), pp.972–7.

Dallmann, R., Viola, A.U., Tarokh, L., Cajochen, C. and Brown, S.A., 2012. The Human Circadian Metabolome. *Proceedings of the National Academy of Sciences*, 109(7), pp.2625–9.

Deininger, M.W., Goldman, J.M., Lydon, N. and Melo, J. V, 1997. The Tyrosine Kinase Inhibitor CGP57148B Selectively Inhibits the Growth of BCR-ABL-Positive Cells. *Blood*, 90(9), pp.3691–8.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G.L., 2006. Greengenes: A Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7), pp.5069–72.

Dettmer, K., Aronov, P.A. and Hammock, B.D., 2007. Mass Spectrometry-Based Metabolomics. *Mass Spectrometry Reviews*, 26(1), pp.51–78.

Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C.R., Yu, W.-H., Lakshmanan, A. and Wade, W.G., 2010. The Human Oral Microbiome. *Journal of Bacteriology*, 192(19), pp.5002–17.

Dickson, R.P., Erb-Downward, J.R. and Huffnagle, G.B., 2013. The Role of the Bacterial Microbiome in Lung Disease. *Expert Review of Respiratory Medicine*, 7(3), pp.245–257.

Dinan, T.G. and Cryan, J.F., 2012. Regulation of the Stress Response by the Gut Microbiota: Implications for Psychoneuroendocrinology. *Psychoneuroendocrinology*, 37(9), pp.1369–78.

Duportet, X., Aggio, R.B.M., Carneiro, S. and Villas-Bôas, S.G., 2011. The Biological Interpretation of Metabolomic Data can be Misled by the Extraction Method Used. *Metabolomics*, 8(3), pp.410–421.

Eisen, J.A. and Wu, M., 2002. Phylogenetic Analysis and Gene Functional Predictions: Phylogenomics in Action. *Theoretical Population Biology*, 61(4), pp.481–487.

Erb-Downward, J.R., Thompson, D.L., Han, M.K., Freeman, C.M., McCloskey, L., Schmidt, L.A., Young, V.B., Toews, G.B., Curtis, J.L., Sundaram, B., Martinez, F.J. and Huffnagle, G.B., 2011. Analysis of the Lung Microbiome in the 'Healthy' Smoker and in COPD. *PLoS One*, 6(2).

Erkiliç, S., Özsaraç, C. and Küllü, S., 2003. Sputum Cytology for the Diagnosis of Lung Cancer. *Acta Cytologica*, 47(6), pp.1023–1027.

Eutamene, H., Lamine, F., Chabo, C., Theodorou, V., Rochat, F., Bergonzelli, G.E., Corthesy-Theulaz, I., Fioramonti, J. and Bueno, L., 2007. Synergy Between *Lactobacillus paracasei* and its Bacterial Products to Counteract Stress-Induced Gut Permeability and Sensitivity Increase in Rats. *Journal of Nutrition*, 137(8), pp.1901–1907.

Favé, G., Beckmann, M., Lloyd, A.J., Zhou, S., Harold, G., Lin, W., Tailliart, K., Xie, L., Draper, J. and Mathers, J.C., 2011. Development and Validation of a Standardized Protocol to Monitor Human Dietary Exposure by Metabolite Fingerprinting of Urine Samples. *Metabolomics*, 7(4), pp.469–484.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C. and Parkin, D.M., 2010. Estimates of Worldwide Burden of Cancer in 2008. *International Journal of Cancer*, 127(12), pp.2893–2917.

Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., van Ommen, B., Lindon, J.C. and Sansone, S.-A., 2007. The Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3), pp.175–178.

Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. and Thurston, M., 2006. Open Software for Biologists: From Famine to Feast. *Nature Biotechnology*, 24(7), pp.801–3.

Field, R.W., Steck, D.J., Smith, B.J., Brus, C.P., Fisher, E.L., Neuberger, J.S., Platz, C.E., Robinson, R.A., Woolson, R.F. and Lynch, C.F., 2000. Residential Radon Gas Exposure and Lung Cancer - The Iowa Radon Lung Cancer Study. *American Journal of Epidemiology*, 151(11), pp.1091–1102.

Foss, K.M., Sima, C., Ugolini, D., Neri, M., Allen, K.E. and Weiss, G.J., 2011. miR-1254 and miR-574-5p: Serum-Based microRNA Biomarkers for Early-Stage Non-Small Cell Lung Cancer. *Journal of Thoracic Oncology*, 6(3), pp.482–8.

Foster, J.A. and McVey Neufeld, K.-A., 2013. Gut-Brain Axis: How the Microbiome Influences Anxiety and Depression. *Trends in Neurosciences*, 36(5), pp.305–12.

Foster, J.S., Wheeler, R.M. and Pamphile, R., 2014. Host-Microbe Interactions in Microgravity: Assessment and Implications. *Life*, 4(2), pp.250–66.

Garcha, D.S., Thurston, S.J., Patel, A.R.C., Mackay, A.J., Goldring, J.J.P., Donaldson, G.C., McHugh, T.D. and Wedzicha, J.A., 2012. Changes in Prevalence and Load of Airway Bacteria Using Quantitative PCR in Stable and Exacerbated COPD. *Thorax*, 67(12), pp.1075–80.

Gevers, D., Pop, M., Schloss, P.D. and Huttenhower, C., 2012. Bioinformatics for the Human Microbiome Project. *PLoS Computational Biology*, 8(11), p.e1002779.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M. and Nelson, K.E., 2006. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science*, 312(5778), pp.1355–9.

Gleeson, M., 2007. Immune Function in Sport and Exercise. *Journal of Applied Physiology*, 103(2), pp.693–9.

Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2014. Global Strategy for the Diagnosis, Management and Prevention of COPD.

Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R. and Ley, R.E., 2014. Conducting a Microbiome Study. *Cell*, 158(2), pp.250–262.

Gottschalk, S., Anderson, N., Hainz, C., Eckhardt, S.G. and Serkova, N.J., 2004. Imatinib (STI571)-Mediated Changes in Glucose Metabolism in Human Leukemia BCR-ABL-Positive Cells. *Clinical Cancer Research*, 10(19), pp.6661–8.

Van der Greef, J., Stroobant, P. and van der Heijden, R., 2004. The Role of Analytical Sciences in Medical Systems Biology. *Current Opinion in Chemical Biology*, 8(5), pp.559–65.

Grice, E.A. and Segre, J.A., 2012. The Human Microbiome: Our Second Genome. *Annual Review of Genomics and Human Genetics*, 13, pp.151–70.

Griffin, J., 2003. Metabonomics: NMR Spectroscopy and Pattern Recognition Analysis of Body Fluids and Tissues for Characterisation of Xenobiotic Toxicity and Disease Diagnosis. *Current Opinion in Chemical Biology*, 7(5), pp.648–654.

Groeneweg, F.L., Karst, H., de Kloet, E.R. and Joëls, M., 2011. Rapid Non-Genomic Effects of Corticosteroids and Their Role in the Central Stress Response. *The Journal of Endocrinology*, 209(2), pp.153–67.

Groenewegen, K.H. and Wouters, E.F.M., 2003. Bacterial Infections in Patients Requiring Admission for an Acute Exacerbation of COPD: A One Year Prospective Study. *Respiratory Medicine*, 97(7), pp.770–777.

Guss, A.M., Roeselers, G., Newton, I.L.G., Young, C.R., Klepac-Ceraj, V., Lory, S. and Cavanaugh, C.M., 2011. Phylogenetic and Metabolic Diversity of Bacteria Associated with Cystic Fibrosis. *ISME Journal*, 5(1), pp.20–29.

Guzmán, L., Depix, M.S., Salinas, A.M., Roldán, R., Aguayo, F., Silva, A. and Vinet, R., 2012. Analysis of Aberrant Methylation on Promoter Sequences of Tumor Suppressor Genes and Total DNA in Sputum Samples: A Promising Tool for Early Detection of COPD and Lung Cancer in Smokers. *Diagnostic Pathology*, 7, p.87.

Halbert, R.J., Natoli, J.L., Gano, A., Badamgarav, E., Buist, A.S. and Mannino, D.M., 2006. Global Burden of COPD: Systematic Review and Meta-Analysis. *The European Respiratory Journal*, 28(3), pp.523–32.

Han, M.K., Huang, Y.J., Lipuma, J.J., Boushey, H.A., Boucher, R.C., Cookson, W.O., Curtis, J.L., Erb-Downward, J., Lynch, S. V, Sethi, S., Toews, G.B., Young, V.B., Wolfgang, M.C., Huffnagle, G.B. and Martinez, F.J., 2012. Significance of the Microbiome in Obstructive Lung Disease. *Thorax*, 67(5), pp.456–63.

Hanahan, D. and Weinberg, R.A., 2000. The Hallmarks of Cancer. *Cell*, 100(1), pp.57–70.

Hanahan, D. and Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5), pp.646–674.

Helicobacter and Cancer Collaborative Group, 2001. Gastric Cancer and *Helicobacter pylori*: A Combined Analysis of 12 Case Control Studies Nested Within Prospective Cohorts. *Gut*, 49(3), pp.347–353.

Heuvers, M.E., Aerts, J.G.J. V, Hegmans, J.P., Veltman, J.D., Uitterlinden, A.G., Ruiter, R., Rodenburg, E.M., Hofman, A., Bakker, M., Hoogsteden, H.C., Stricker, B.H. and van Klaveren, R.J., 2012. History of Tuberculosis as an Independent Prognostic Factor for Lung Cancer Survival. *Lung Cancer*, 76(3), pp.452–6.

Hilty, M., Burke, C., Pedro, H., Cardenas, P., Bush, A., Bossley, C., Davies, J., Ervine, A., Poulter, L., Pachter, L., Moffatt, M.F. and Cookson, W.O.C., 2010. Disordered Microbial Communities in Asthmatic Airways. *PLoS One*, 5(1).

Hirschmann, J. V, 2000. Do Bacteria Cause Exacerbations of COPD? *Chest*, 118(1), pp.193–203.

Hongoh, Y., Ohkuma, M. and Kudo, T., 2003. Molecular Analysis of Bacterial Microbiota in the Gut of the Termite *Reticulitermes speratus* (Isoptera: Rhinotermitidae). *FEMS Microbiology Ecology*, 44(2), pp.231–42.

Hori, S., Nishiumi, S., Kobayashi, K., Shinohara, M., Hatakeyama, Y., Kotani, Y., Hatano, N., Maniwa, Y., Nishio, W., Bamba, T., Fukusaki, E., Azuma, T., Takenawa, T., Nishimura, Y. and Yoshida, M., 2011. A Metabolomic Approach to Lung Cancer. *Lung Cancer*, 74(2), pp.284–92.

Hosgood, H.D., Sapkota, A.R., Rothman, N., Rohan, T., Hu, W., Xu, J., Vermeulen, R., He, X., White, J.R., Wu, G., Wei, F., Mongodin, E.F. and Lan, Q., 2014. The Potential Role of Lung Microbiota in Lung Cancer Attributed to Household Coal Burning Exposures. *Environmental and Molecular Mutagenesis*.

Huang, Y.J., Kim, E., Cox, M.J., Brodie, E.L., Brown, R., Wiener-Kronish, J.P. and Lynch, S. V, 2010. A Persistent and Diverse Airway Microbiota Present During Chronic Obstructive Pulmonary Disease Exacerbations. *OMICS: A Journal of Integrative Biology*, 14(1), pp.9–59.

Huang, Y.J., Nelson, C.E., Brodie, E.L., DeSantis, T.Z., Baek, M.S., Liu, J., Woyke, T., Allgaier, M., Bristow, J., Wiener-Kronish, J.P., Sutherland, E.R., King, T.S., Icitovic, N., Martin, R.J., Calhoun, W.J.,

Castro, M., Denlinger, L.C., DiMango, E., Kraft, M., Peters, S.P., Wasserman, S.I., Wechsler, M.E., Boushey, H.A., Lynch, S. V, Natl Heart, L. and Blood Inst, A., 2011. Airway Microbiota and Bronchial Hyperresponsiveness in Patients with Suboptimally Controlled Asthma. *Journal of Allergy and Clinical Immunology*, 127(2), pp.372–689.

Hubers, A.J., Prinsen, C.F.M., Sozzi, G., Witte, B.I. and Thunnissen, E., 2013. Molecular Sputum Analysis for the Diagnosis of Lung Cancer. *British Journal of Cancer*, 109(3), pp.530–7.

Huffnagle, G.B. and Noverr, M.C., 2013. The Emerging World of the Fungal Microbiome. *Trends in Microbiology*, 21(7), pp.334–41.

Humphrey, S.P. and Williamson, R.T., 2001. A Review of Saliva: Normal Composition, Flow, and Function. *The Journal of Prosthetic Dentistry*, 85(2), pp.162–9.

Husgafvel-Pursiainen, K., 2004. Genotoxicity of Environmental Tobacco Smoke: A Review. *Mutation Research*, 567(2-3), pp.427–45.

Infante, M., Cavuto, S., Lutman, F.R., Brambilla, G., Chiesa, G., Ceresoli, G., Passera, E., Angeli, E., Chiarenza, M., Aranzulla, G., Cariboni, U., Errico, V., Inzirillo, F., Bottoni, E., Voulaz, E., Alloisio, M., Destro, A., Roncalli, M., Santoro, A., Ravasi, G. and Dante Study, G., 2009. A Randomized Study of Lung Cancer Screening with Spiral Computed Tomography - Three-Year Results from the DANTE Trial. *American Journal of Respiratory and Critical Care Medicine*, 180(5), pp.445–453.

Isitmangil, T., Isitmangil, G., Budak, Y., Aydilek, R. and Celenk, M., 2001. Comparison of Serum and Bronchoalveolar Lavage Fluid Sialic Acid Levels Between Malignant and Benign Lung Diseases. *BMC Pulmonary Medicine*, 1(1), p.4.

Issaq, H.J., Nativ, O., Waybright, T., Luke, B., Veenstra, T.D., Issaq, E.J., Kravstov, A. and Mullerad, M., 2008. Detection of Bladder Cancer in Human Urine by Metabolomic Profiling using High Performance Liquid Chromatography Mass Spectrometry. *The Journal of Urology*, 179(6), pp.2422–6.

Jarvis, R.M., Broadhurst, D., Johnson, H., O'Boyle, N.M. and Goodacre, R., 2006. PYCHEM: A Multivariate Analysis Package for Python. *Bioinformatics*, 22(20), pp.2565–6.

Jemal, A., Center, M.M., DeSantis, C. and Ward, E.M., 2010a. Global Patterns of Cancer Incidence and Mortality Rates and Trends. *Cancer Epidemiology Biomarkers and Prevention*, 19(8), pp.1893–1907.

Jemal, A., Siegel, R., Xu, J.Q. and Ward, E., 2010b. Cancer Statistics, 2010. *CA: A Cancer Journal for Clinicians*, 60(5), pp.277–300.

Jenkinson, H.F. and Lamont, R.J., 2005. Oral Microbial Communities in Sickness and in Health. *Trends in Microbiology*, 13(12), pp.589–95.

Jones, A.W., Cameron, S.J.S., Thatcher, R., Beecroft, M.S., Mur, L.A.J. and Davison, G., 2014. Effects of Bovine Colostrum Supplementation on Upper Respiratory Illness in Active Males. *Brain, Behavior, and Immunity*, 39, pp.194–203.

Jones, P.W., 2009. Health Status and the Spiral of Decline. *COPD*, 6(1), pp.59–63.

Jones, R. and Golding, J., 2009. Choosing the Types of Biological Sample to Collect in Longitudinal Birth Cohort Studies. *Paediatric and Perinatal Epidemiology*, 23 Suppl 1, pp.103–13.

Karouia, F., Santos, O., Valdivia-Silva, J.E., Jones, J., Greenberger, J.S. and Epperly, M.W., 2014. Impact of Whole Body Irradiation on the Intestinal Microbiome - Considerations for Space Flight. *40th COSPAR Scientific Assembly. Held 2-10 August 2014*.

Khalaila, R., Cohen, M. and Zidan, J., 2014. Is Salivary pH a Marker of Depression Among Older Spousal Caregivers for Cancer Patients? *Behavioral Medicine*, 40(2), pp.71–80.

Khan, A.A., Shrivastava, A. and Khurshid, M., 2012. Normal to Cancer Microbiome Transformation and its Implication in Cancer Diagnosis. *Biochimica et Biophysica Acta*, 1826(2), pp.331–7.

Khoo, A.-L., Koenen, H.J.P.M., Chai, L.Y.A., Sweep, F.C.G.J., Netea, M.G., van der Ven, A.J.A.M. and Joosten, I., 2012. Seasonal Variation in Vitamin D_3 Levels is Paralleled by Changes in the Peripheral Blood Human T Cell Compartment. *PLoS One*, 7(1), p.e29250.

Kind, T., Tolstikov, V., Fiehn, O. and Weiss, R.H., 2007. A Comprehensive Urinary Metabolomic Approach for Identifying Kidney Cancer. *Analytical Biochemistry*, 363(2), pp.185–95.

Kinross, J.M., Darzi, A.W. and Nicholson, J.K., 2011. Gut Microbiome-Host Interactions in Health and Disease. *Genome Medicine*, 3(3), p.14.

Kleiner, S.M., 1999. Water: An Essential But Overlooked Nutrient. *Journal of the American Dietetic Association*, 99(2), pp.200–6.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glöckner, F.O., 2013. Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and Next-Generation Sequencing-Based Diversity Studies. *Nucleic Acids Research*, 41(1), p.e1.

Klink, M., Bednarska, K., Blus, E., Kielbik, M. and Sulowska, Z., 2012. Seasonal Changes in Activities of Human Neutrophils *in vitro*. *Inflammation Research*, 61(1), pp.11–6.

Van der Kloet, F.M., Tempels, F.W.A., Ismail, N., van der Heijden, R., Kasper, P.T., Rojas-Cherto, M., van Doorn, R., Spijksma, G., Koek, M., van der Greef, J., Mäkinen, V.P., Forsblom, C., Holthöfer, H., Groop, P.H., Reijmers, T.H. and Hankemeier, T., 2012. Discovery of Early-Stage Biomarkers for Diabetic Kidney Disease using Mass Spectrometry Based Metabolomics (The FinnDiane Study). *Metabolomics*, 8(1), pp.109–119.

Kodama, Y., Shumway, M. and Leinonen, R., 2012. The Sequence Read Archive: Explosive Growth of Sequencing Data. *Nucleic Acids Research*, 40(Database issue), pp.D54–6.

Konturek, P.C., Brzozowski, T. and Konturek, S.J., 2011. Stress and the Gut: Pathophysiology, Clinical Consequences, Diagnostic Approach and Treatment Options. *Journal of Physiology and Pharmacology*, 62(6), pp.591–9.

Koolhaas, J.M., Bartolomucci, A., Buwalda, B., de Boer, S.F., Flügge, G., Korte, S.M., Meerlo, P., Murison, R., Olivier, B., Palanza, P., Richter-Levin, G., Sgoifo, A., Steimer, T., Stiedl, O., van Dijk, G., Wöhr, M. and Fuchs, E., 2011. Stress Revisited: A Critical Evaluation of the Stress Concept. *Neuroscience and Biobehavioral Reviews*, 35(5), pp.1291–301.

Korbel, D.S., Schneider, B.E. and Schaible, U.E., 2008. Innate Immunity in Tuberculosis: Myths and Truth. *Microbes and Infection*, 10(9), pp.995–1004.

Koshiol, J., Flores, R., Lam, T.K., Taylor, P.R., Weinstein, S.J., Virtamo, J., Albanes, D., Perez-Perez, G., Caporaso, N.E. and Blaser, M.J., 2012. *Helicobacter pylori* Seropositivity and Risk of Lung Cancer. *PLoS One*, 7(2), p.e32106.

Kosuge, T., Mashima, J., Kodama, Y., Fujisawa, T., Kaminuma, E., Ogasawara, O., Okubo, K., Takagi, T. and Nakamura, Y., 2014. DDBJ Progress Report: A New Submission System for Leading to a Correct Annotation. *Nucleic Acids Research*, 42(Database issue), pp.D44–9.

Kote-Jarai, Z., Easton, D.F., Stanford, J.L., Ostrander, E.A., Schleutker, J., Ingles, S.A., Schaid, D., Thibodeau, S., Dörk, T., Neal, D., Donovan, J., Hamdy, F., Cox, A., Maier, C., Vogel, W., Guy, M., Muir, K., Lophatananon, A., Kedda, M.-A., Spurdle, A., Steginga, S., John, E.M., Giles, G., Hopper, J., Chappuis, P.O., Hutter, P., Foulkes, W.D., Hamel, N., Salinas, C.A., Koopmeiners, J.S., Karyadi, D.M., Johanneson, B., Wahlfors, T., Tammela, T.L., Stern, M.C., Corral, R., McDonnell, S.K., Schürmann, P., Meyer, A., Kuefer, R., Leongamornlert, D.A., Tymrakiewicz, M., Liu, J.-F., O'Mara, T., Gardiner, R.A.F., Aitken, J., Joshi, A.D., Severi, G., English, D.R., Southey, M., Edwards, S.M., Al Olama, A.A. and Eeles, R.A., 2008. Multiple Novel Prostate Cancer Predisposition Loci Confirmed by an International Study: The PRACTICAL Consortium. *Cancer Epidemiology, Biomarkers and Prevention*, 17(8), pp.2052–61.

Kuczynski, J., Costello, E.K., Nemergut, D.R., Zaneveld, J., Lauber, C.L., Knights, D., Koren, O., Fierer, N., Kelley, S.T., Ley, R.E., Gordon, J.I. and Knight, R., 2010. Direct Sequencing of the Human Microbiome Readily Reveals Community Differences. *Genome Biology*, 11(5).

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D. and Knight, R., 2012. Experimental and Analytical Tools for Studying the Human Microbiome. *Nature Reviews: Genetics*, 13(1), pp.47–58.

Laroumagne, S., Lepage, B., Hermant, C., Plat, G., Phelippeau, M., Bigay-Game, L., Lozano, S., Guibert, N., Segonds, C., Mallard, V., Augustin, N., Didier, A. and Mazieres, J., 2013. Bronchial Colonisation in Patients with Lung Cancer: A Prospective Study. *The European Respiratory Journal*, 42(1), pp.220–9.

Laurence, M., Hatzis, C. and Brash, D.E., 2014. Common Contaminants in Next-Generation Sequencing that Hinder Discovery of Low-Abundance Microbes. *PLoS ONE*, 9(5), p.e97876.

Lawton, K.A., Berger, A., Mitchell, M., Milgram, K.E., Evans, A.M., Guo, L., Hanson, R.W., Kalhan, S.C., Ryals, J.A. and Milburn, M. V, 2008. Analysis of the Adult Human Plasma Metabolome. *Pharmacogenomics*, 9(4), pp.383–97.

Laxman, B., Morris, D.S., Yu, J., Siddiqui, J., Cao, J., Mehra, R., Lonigro, R.J., Tsodikov, A., Wei, J.T., Tomlins, S.A. and Chinnaiyan, A.M., 2008. A First-Generation Multiplex Biomarker Analysis of Urine for the Early Detection of Prostate Cancer. *Cancer Research*, 68(3), pp.645–9.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V. and Cochrane, G., 2011. The European Nucleotide Archive. *Nucleic Acids Research*, 39(Database issue), pp.D28–31.

Lewis, P.D., Lewis, K.E., Ghosal, R., Bayliss, S., Lloyd, A.J., Wills, J., Godfrey, R., Kloer, P. and Mur, L.A.J., 2010. Evaluation of FTIR Spectroscopy as a Diagnostic Tool for Lung Cancer Using Sputum. *BMC Cancer*, 10.

Li, K., Bihan, M., Yooseph, S. and Methé, B.A., 2012. Analyses of the Microbial Diversity across the Human Microbiome. *PLoS ONE*, 7(6), p.e32118.

Li, Y., Hao, F., Fang, F., Zhang, L., Zheng, H., Ma, R.Z., Wang, X.Y., Xu, H.J. and Zang, Y.X., 2013. Analysis of Flora Distribution and Drug Resistance in Sputum Culture from Patients with Lung Cancer. *Advanced Materials Research*, 641-642, pp.625–629.

Lim, Y.W., Evangelista, J.S., Schmieder, R., Bailey, B., Haynes, M., Furlan, M., Maughan, H., Edwards, R., Rohwer, F. and Conrad, D., 2014. Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis. *Journal of Clinical Microbiology*, 52(2), pp.425–37.

Liu, B., Faller, L.L., Klitgord, N., Mazumdar, V., Ghodsi, M., Sommer, D.D., Gibbons, T.R., Treangen, T.J., Chang, Y.-C., Li, S., Stine, O.C., Hasturk, H., Kasif, S., Segrè, D., Pop, M. and Amar, S., 2012. Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. *PLoS ONE*, 7(6), p.e37919.

Loman, N.J., Misra, R. V, Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. and Pallen, M.J., 2012. Performance Comparison of Benchtop High-Throughput Sequencing Platforms. *Nature Biotechnology*, 30(5), pp.434–9.

Lopez, A.D., Collishaw, N.E. and Piha, T., 1994. A Descriptive Model of the Cigarette Epidemic in Developed Countires. *Tobacco Control*, 3, pp.242–247.

Lopez, A.D., Shibuya, K., Rao, C., Mathers, C.D., Hansell, A.L., Held, L.S., Schmid, V. and Buist, S., 2006. Chronic Obstructive Pulmonary Disease: Current Burden and Future Projections. *The European Respiratory Journal*, 27(2), pp.397–412.

Losos, J.B., Arnold, S.J., Bejerano, G., Brodie, E.D., Hibbett, D., Hoekstra, H.E., Mindell, D.P., Monteiro, A., Moritz, C., Orr, H.A., Petrov, D.A., Renner, S.S., Ricklefs, R.E., Soltis, P.S. and Turner, T.L., 2013. Evolutionary Biology for the 21st Century. *PLoS Biology*, 11(1), p.e1001466.

Lower, R., Lower, J. and Kurth, R., 1996. The Viruses in All of Us: Characteristics and Biological Significance of Human Endogenous Retrovirus Sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11), pp.5177–5184.

Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K. and Knight, R., 2012. Diversity, Stability and Resilience of the Human Gut Microbiota. *Nature*, 489(7415), pp.220–30.

Madsen, R., Lundstedt, T. and Trygg, J., 2010. Chemometrics in Metabolomics: A Review in Human Disease Diagnosis. *Analytica Chimica Acta*, 659(1-2), pp.23–33.

Maeda, H., Fujimoto, C., Haruki, Y., Maeda, T., Kokeguchi, S., Petelin, M., Arai, H., Tanimoto, I., Nishimura, F. and Takashiba, S., 2003. Quantitative Real-Time PCR Using TaqMan and SYBR Green for *Actinobacillus actinomycetemcomitans*, *Porphyromonas gingivalis*, *Prevotella intermedia*, *tetQ* Gene and Total Bacteria. *FEMS Immunology and Medical Microbiology*, 39(1), pp.81–86.

Magoč, T. and Salzberg, S.L., 2011. FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. *Bioinformatics*, 27(21), pp.2957–63.

Mahenthiralingam, E., 2014. Emerging Cystic Fibrosis Pathogens and the Microbiome. *Paediatric Respiratory Reviews*, 15 Supplem, pp.13–5.

Mamas, M., Dunn, W.B., Neyses, L. and Goodacre, R., 2011. The Role of Metabolites and Metabolomics in Clinically Applicable Biomarkers of Disease. *Archives of Toxicology*, 85(1), pp.5–17.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., Roca, J. and Dore, J., 2006. Reduced Diversity of Faecal Microbiota in Crohn's Disease Revealed by a Metagenomic Approach. *Gut*, 55(2), pp.205–11.

Mannino, D.M. and Buist, A.S., 2007. Global Burden of COPD: Risk Factors, Prevalence, and Future Trends. *Lancet*, 370(9589), pp.765–73.

Mantilla Gomez, S., Danser, M.M., Sipos, P.M., Rowshani, B., van der Velden, U. and van der Weijden, G.A., 2001. Tongue Coating and Salivary Bacterial Counts in Healthy/Gingivitis Subjects and Periodontitis Patients. *Journal of Clinical Periodontology*, 28(10), pp.970–978.

Marchesi, J.R., Dutilh, B.E., Hall, N., Peters, W.H.M., Roelofs, R., Boleij, A. and Tjalsma, H., 2011. Towards the Human Colorectal Cancer Microbiome. *PLoS One*, 6(5), p.e20447.

Mardanov, A. V, Babykin, M.M., Beletsky, A. V, Grigoriev, A.I., Zinchenko, V. V, Kadnikov, V. V, Kirpichnikov, M.P., Mazur, A.M., Nedoluzhko, A. V, Novikova, N.D., Prokhortchouk, E.B., Ravin, N. V, Skryabin, K.G. and Shestakov, S. V, 2013. Metagenomic Analysis of the Dynamic Changes in the Gut Microbiome of the Participants of the MARS-500 Experiment: Simulating Long Term Space Flight. *Acta Naturae*, 5(3), pp.116–25.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X. V, Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M., 2005. Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature*, 437(7057), pp.376–80.

Marion, C., Burnaugh, A.M., Woodiga, S.A. and King, S.J., 2011. Sialic Acid Transport Contributes to Pneumococcal Colonization. *Infection and Immunity*, 79(3), pp.1262–9.

Markle, J.G.M., Frank, D.N., Mortin-Toth, S., Robertson, C.E., Feazel, L.M., Rolle-Kampczyk, U., von Bergen, M., McCoy, K.D., Macpherson, A.J. and Danska, J.S., 2013. Sex Differences in the Gut Microbiome Drive Hormone-Dependent Regulation of Autoimmunity. *Science*, 339(6123), pp.1084–8.

Mathé, E.A., Patterson, A.D., Haznadar, M., Manna, S.K., Krausz, K.W., Bowman, E.D., Shields, P.G., Idle, J.R., Smith, P.B., Anami, K., Kazandjian, D.G., Hatzakis, E., Gonzalez, F.J. and Harris, C.C., 2014. Noninvasive Urinary Metabolomic Profiling Identifies Diagnostic and Prognostic Markers in Lung Cancer. *Cancer Research*, 74(12), pp.3259–70.

Mathers, C.D. and Loncar, D., 2006. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*, 3(11), p.e442.

Matias, I., Gatta-Cherifi, B., Tabarin, A., Clark, S., Leste-Lasserre, T., Marsicano, G., Piazza, P.V. and Cota, D., 2012. Endocannabinoids Measurement in Human Saliva as Potential Biomarker of Obesity. *PLoS One*, 7(7), p.e42399.

McGuire, A.L., Colgrove, J., Whitney, S.N., Diaz, C.M., Bustillos, D. and Versalovic, J., 2008. Ethical, Legal, and Social Considerations in Conducting the Human Microbiome Project. *Genome Research*, 18(12), pp.1861–1864.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. and Edwards, R.A., 2008. The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. *BMC Bioinformatics*, 9(1), p.386.

Monso, E., Garcia-Aymerich, J., Soler, N., Farrero, E., Felez, M.A., Anto, J.M. and Torres, A., 2003. Bacterial Infection in Exacerbated COPD with Changes in Sputum Characteristics. *Epidemiology and Infection*, 131(1), pp.799–804.

Morgan, F., Barker, G., Briggs, C., Price, R. and Keys, H., 2007. *Environmental Domains of Antarctica Version 2.0 Final Report*.

Morris, A., Beck, J.M., Schloss, P.D., Campbell, T.B., Crothers, K., Curtis, J.L., Flores, S.C., Fontenot, A.P., Ghedin, E., Huang, L., Jablonski, K., Kleerup, E., Lynch, S. V, Sodergren, E., Twigg, H., Young, V.B., Bassis, C.M., Venkataraman, A., Schmidt, T.M. and Weinstock, G.M., 2013. Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers. *American Journal of Respiratory and Critical Care Medicine*, 187(10), pp.1067–75.

Muñoz-Almagro, C., Gala, S., Selva, L., Jordan, I., Tarragó, D. and Pallares, R., 2011. DNA Bacterial Load in Children and Adolescents with Pneumococcal Pneumonia and Empyema. *European Journal of Clinical Microbiology and Infectious Diseases*, 30(3), pp.327–35.

Narikiyo, M., Tanabe, C., Yamada, Y., Igaki, H., Tachimori, Y., Kato, H., Muto, M., Montesano, R., Sakamoto, H., Nakajima, Y. and Sasaki, H., 2004. Frequent and Preferential Infection of *Treponema denticola*, *Streptococcus mitis*, and *Streptococcus anginosus* in Esophageal Cancers. *Cancer Science*, 95(7), pp.569–574.

Nasidze, I., Li, J., Quinque, D., Tang, K. and Stoneking, M., 2009. Global Diversity in the Human Salivary Microbiome. *Genome Research*, 19(4), pp.636–43.

Nawa, T., Nakagawa, T., Mizoue, T., Kusano, S., Chonan, T., Fukai, S. and Endo, K., 2012. Long-Term Prognosis of Patients with Lung Cancer Detected on Low-Dose Chest Computed Tomography Screening. *Lung Cancer*, 75(2), pp.197–202.

Neyraud, E., Tremblay-Franco, M., Gregoire, S., Berdeaux, O. and Canlet, C., 2012. Relationships Between the Metabolome and the Fatty Acid Composition of Human Saliva: Effects of Stimulation. *Metabolomics*, 9(1), pp.213–222.

Nowotarski, S.L., Woster, P.M. and Casero, R.A., 2013. Polyamines and Cancer: Implications for Chemotherapy and Chemoprevention. *Expert Reviews in Molecular Medicine*, 15, p.e3.

Nseir, S., Di Pompeo, C., Cavestri, B., Jozefowicz, E., Nyunga, M., Soubrier, S., Roussel-Delvallez, M., Saulnier, F., Mathieu, D. and Durocher, A., 2006. Multiple-Drug-Resistant Bacteria in Patients with Severe Acute Exacerbation of Chronic Obstructive Pulmonary Disease: Prevalence, Risk Factors, and Outcome. *Critical Care Medicine*, 34(12), pp.2959–2966.

Ohigashi, S., Sudo, K., Kobayashi, D., Takahashi, O., Takahashi, T., Asahara, T., Nomoto, K. and Onodera, H., 2013. Changes of the Intestinal Microbiota, Short Chain Fatty Acids, and Fecal pH in Patients with Colorectal Cancer. *Digestive Diseases and Sciences*, 58(6), pp.1717–26.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. and Vonstein, V., 2005. The Subsystems Approach to Genome Annotation and its use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, 33(17), pp.5691–702.

Parr, D.G., White, A.J., Bayley, D.L., Guest, P.J. and Stockley, R.A., 2006. Inflammation in Sputum Relates to Progression of Disease in Subjects with COPD: A Prospective Descriptive Study. *Respiratory Research*, 7(1), p.136.

Patel, I.S., 2002. Relationship Between Bacterial Colonisation and the Frequency, Character, and Severity of COPD Exacerbations. *Thorax*, 57(9), pp.759–764.

Paz-Elizur, T., Krupsky, M., Blumenstein, S., Elinger, D., Schechtman, E. and Livneh, Z., 2003. DNA Repair Activity for Oxidative Damage and Risk of Lung Cancer. *Journal of the National Cancer Institute*, 95(17), pp.1312–1319.

Pearson, H., 2007. Meet the Human Metabolome. *Nature*, 446(7131), p.8.

Perkins, A., Osorio, S., Serrano, M., del Ray, M.C., Sarria, C., Domingo, D. and Lopez-Brea, M., 2003. A Case of Endocarditis Due to *Granulicatella adiacens*. *Clinical Microbiology and Infection*, 9(6), pp.576–577.

Peters, B.M., Shirtliff, M.E. and Jabra-Rizk, M.A., 2010. Antimicrobial Peptides: Primeval Molecules or Future Drugs? *PLoS Pathogens*, 6(10), p.e1001067.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R.,

Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B. and Guyer, M., 2009. The NIH Human Microbiome Project. *Genome Research*, 19(12), pp.2317–23.

Pfaffe, T., Cooper-White, J., Beyerlein, P., Kostner, K. and Punyadeera, C., 2011. Diagnostic Potential of Saliva: Current State and Future Applications. *Clinical Chemistry*, 57(5), pp.675–87.

Pragman, A.A., Kim, H.B., Reilly, C.S., Wendt, C. and Isaacson, R.E., 2012. The Lung Microbiome in Moderate and Severe Chronic Obstructive Pulmonary Disease. *PLoS One*, 7(10), p.e47305.

Prakash, T. and Taylor, T.D., 2012. Functional Assignment of Metagenomic Data: Challenges and Applications. *Briefings in Bioinformatics*, 13(6), pp.711–727.

Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., Loomer, P., Armitage, G.C. and Relman, D.A., 2012. Evidence of a Robust Resident Bacteriophage Population Revealed Through Analysis of the Human Salivary Virome. *The ISME Journal*, 6(5), pp.915–26.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D. and Wang, J., 2010. A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature*, 464(7285), pp.59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich,

S.D., Nielsen, R., Pedersen, O., Kristiansen, K. and Wang, J., 2012. A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes. *Nature*, 490(7418), pp.55–60.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O., 2013. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Research*, 41(Database Issue), pp.D590–6.

Relman, D.A. and Falkow, S., 2001. The Meaning and Impact of the Human Genome Sequence for Microbiology. *Trends in Microbiology*, 9(5), pp.206–208.

Renom, F., Yáñez, A., Garau, M., Rubí, M., Centeno, M.-J., Gorriz, M.-T., Medinas, M., Ramis, F., Soriano, J.B. and Alvar Agustí, 2010. Prognosis of COPD Patients Requiring Frequent Hospitalization: Role of Airway Infection. *Respiratory Medicine*, 104(6), pp.840–8.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P. and Woyke, T., 2013. Insights into the Phylogeny and Coding Potential of Microbial Dark Matter. *Nature*, 499(7459), pp.431–7.

Ritz, P. and Berrut, G., 2005. The Importance of Good Hydration for Day-to-Day Health. *Nutrition Reviews*, 63(S1), pp.S6–S13.

Rivera, M.P., Mehta, A.C. and Wahidi, M.M., 2013. Establishing the Diagnosis of Lung Cancer: Diagnosis and Management of Lung Cancer. *Chest*, 143(5 Suppl), p.e142S–65S.

Robertson, D.G., Watkins, P.B. and Reily, M.D., 2011. Metabolomics in Toxicology: Preclinical and Clinical Applications. *Toxicological Sciences*, 120 Supple(1), pp.S146–70.

Rogers, G.B., Carroll, M.P., Serisier, D.J., Hockey, P.M., Jones, G. and Bruce, K.D., 2004. Characterization of Bacterial Community Diversity in Cystic Fibrosis Lung Infections by use of 16S Ribosomal DNA Terminal Restriction Fragment Length Polymorphism Profiling. *Journal of Clinical Microbiology*, 42(11), pp.5176–83.

Rowshani, B., Timmerman, M.F. and Van der Velden, U., 2004. Plaque Development in Relation to the Periodontal Condition and Bacterial Load of the Saliva. *Journal of Clinical Periodontology*, 31(3), pp.214–8.

Saei, A.A. and Barzegari, A., 2012. The Microbiome: The Forgotten Organ of the Astronaut's Body - Probiotics Beyond Terrestrial Limits. *Future Microbiology*, 7(9), pp.1037–46.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. and Walker, A.W., 2014. Reagent and Laboratory Contamination can Critically Impact Sequence-Based Microbiome Analyses. *BMC Biology*, 12(1), p.87.

Salvi, S. and Barnes, P.J., 2010. Is Exposure to Biomass Smoke the Biggest Risk Factor for COPD Globally? *Chest*, 138(1), pp.3–6.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), pp.5463–5467.

Sasaki, H., Ishizuka, T., Muto, M., Nezu, M., Nakanishi, Y., Inagaki, Y., Watanabe, H. and Terada, M., 1998. Presence of *Streptococcus anginosus* DNA in Esophageal Cancer, Dysplasia of Esophagus, and Gastric Cancer. *Cancer Research*, 58(14), pp.2991–5.

Saude, E.J. and Sykes, B.D., 2007. Urine Stability for Metabolomic Studies: Effects of Preparation and Storage. *Metabolomics*, 3(1), pp.19–27.

Scannapieco, F.A., 1994. Saliva-Bacterium Interactions in Oral Microbial Ecology. *Critical Reviews in Oral Biology and Medicine*, 5(3), pp.203–248.

Scannapieco, F.A., 2013. The Oral Microbiome: Its Role in Health and in Oral and Systemic Infections. *Clinical Microbiology Newsletter*, 35(20), pp.163–169.

Schloss, P.D. and Handelsman, J., 2005. Metagenomics for Studying Unculturable Microorganisms: Cutting the Gordian Knot. *Genome Biology*, 6(8), p.229.

Schmidt, M.A. and Goodwin, T.J., 2013. Personalized Medicine in Human Space Flight: Using Omics Based Analyses to Develop Individualized Countermeasures that Enhance Astronaut Safety and Performance. *Metabolomics*, 9(6), pp.1134–1156.

Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C. and Izard, J., 2012. Composition of the Adult Digestive Tract Bacterial Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples. *Genome Biology*, 13(6), p.R42.

Sellitto, M., Bai, G., Serena, G., Fricke, W.F., Sturgeon, C., Gajer, P., White, J.R., Koenig, S.S.K., Sakamoto, J., Boothe, D., Gicquelais, R., Kryszak, D., Puppa, E., Catassi, C., Ravel, J. and Fasano, A., 2012. Proof of Concept of Microbiome-Metabolome Analysis and Delayed Gluten Exposure on Celiac Disease Autoimmunity in Genetically At-Risk Infants. *PLoS One*, 7(3), p.e33387.

Serino, M., 2012. Intestinal MicrobiOMICS to Define Health and Disease in Human and Mice. *Current Pharmaceutical Biotechnology*, 13(5), pp.746–758.

Severi, E., Hood, D.W. and Thomas, G.H., 2007. Sialic Acid Utilization by Bacterial Pathogens. *Microbiology*, 153(Pt 9), pp.2817–22.

Shaughnessy, J., Lewis, L.A., Jarva, H. and Ram, S., 2009. Functional Comparison of the Binding of Factor H Short Consensus Repeat 6 (SCR 6) to Factor H Binding Protein from *Neisseria meningitidis* and the Binding of Factor H SCR 18 to 20 to *Neisseria gonorrhoeae* Porin. *Infection and Immunity*, 77(5), pp.2094–103.

Shen, J., Todd, N.W., Zhang, H., Yu, L., Lingxiao, X., Mei, Y., Guarnera, M., Liao, J., Chou, A., Lu, C.L., Jiang, Z., Fang, H., Katz, R.L. and Jiang, F., 2011. Plasma microRNAs as Potential Biomarkers for Non-Small-Cell Lung Cancer. *Laboratory Investigation*, 91(4), pp.579–87.

Shendure, J. and Ji, H., 2008. Next-Generation DNA Sequencing. *Nature Biotechnology*, 26(10), pp.1135–45.

Shiels, M.S., Albanes, D., Virtamo, J. and Engels, E.A., 2011. Increased Risk of Lung Cancer in Men with Tuberculosis in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study. *Cancer Epidemiology, Biomarkers and Prevention*, 20(4), pp.672–8.

Shiga, K., Tateda, M., Saijo, S., Hori, T., Sato, I., Tateno, H., Matsuura, K., Takasaka, T. and Miyagi, T., 2001. Presence of Streptococcus Infection in Extra-Oropharyngeal Head and Neck Squamous Cell Carcinoma and its Implication in Carcinogenesis. *Oncology Reports*, 8(2), pp.245–8.

Silva, C.L., Passos, M. and Câmara, J.S., 2011. Investigation of Urinary Volatile Organic Metabolites as Potential Cancer Biomarkers by Solid-Phase Microextraction in Combination with Gas Chromatography-Mass Spectrometry. *British Journal of Cancer*, 105(12), pp.1894–904.

Slupsky, C.M., Steed, H., Wells, T.H., Dabbs, K., Schepansky, A., Capstick, V., Faught, W. and Sawyer, M.B., 2010. Urine Metabolite Analysis Offers Potential Early Diagnosis of Ovarian and Breast Cancers. *Clinical Cancer Research*, 16(23), pp.5835–41.

Smith, R.A., Cokkinides, V., Brooks, D., Saslow, D. and Brawley, O.W., 2010. Cancer Screening in the United States, 2010: A Review of Current American Cancer Society Guidelines and Issues in Cancer Screening. *CA: A Cancer Journal for Clinicians*, 60(2), pp.99–119.

Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J.P., Letulle, S., Langella, P., Corthier, G., Tran Van Nhieu, J. and Furet, J.P., 2011. Microbial Dysbiosis in Colorectal Cancer (CRC) Patients. *PLoS One*, 6(1), p.e16393.

Sonnenburg, J.L. and Fischbach, M.A., 2011. Community Health Care: Therapeutic Opportunities in the Human Microbiome. *Science Translational Medicine*, 3(78).

Spratlin, J.L., Serkova, N.J. and Eckhardt, S.G., 2009. Clinical Applications of Metabolomics in Oncology: A Review. *Clinical Cancer Research*, 15(2), pp.431–40.

Stahringer, S.S., Clemente, J.C., Corley, R.P., Hewitt, J., Knights, D., Walters, W.A., Knight, R. and Krauter, K.S., 2012. Nurture Trumps Nature in a Longitudinal Survey of Salivary Bacterial Communities in Twins from Early Adolescence to Early Adulthood. *Genome Research*, 22(11), pp.2146–52.

Stanton, S.J., Mullette-Gillman, O.A. and Huettel, S.A., 2011. Seasonal Variation of Salivary Testosterone in Men, Normally Cycling Women, and Women Using Hormonal Contraceptives. *Physiology and Behaviour*, 104(5), pp.804–8.

Stencel-Baerenwald, J.E., Reiss, K., Reiter, D.M., Stehle, T. and Dermody, T.S., 2014. The Sweet Spot: Defining Virus-Sialic Acid Interactions. *Nature Reviews: Microbiology*, 12(11), pp.739–749.

Sugimoto, M., Saruta, J., Matsuki, C., To, M., Onuma, H., Kaneko, M., Soga, T., Tomita, M. and Tsukinoki, K., 2012. Physiological and Environmental Parameters Associated with Mass Spectrometry-Based Salivary Metabolomic Profiles. *Metabolomics*, 9(2), pp.454–463.

Sugimoto, M., Wong, D.T., Hirayama, A., Soga, T. and Tomita, M., 2010. Capillary Electrophoresis Mass Spectrometry-Based Saliva Metabolomics Identified Oral, Breast and Pancreatic Cancer-Specific Profiles. *Metabolomics*, 6(1), pp.78–95.

Sze, M.A., Dimitriu, P.A., Hayashi, S., Elliott, W.M., McDonough, J.E., Gosselink, J. V, Cooper, J., Sin, D.D., Mohn, W.W. and Hogg, J.C., 2012. The Lung Tissue Microbiome in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, 185(10), pp.1073–80.

Takahashi, N., Washio, J. and Mayanagi, G., 2010. Metabolomics of Supragingival Plaque and Oral Bacteria. *Journal of Dental Research*, 89(12), pp.1383–8.

Takeda, I., Stretch, C., Barnaby, P., Bhatnager, K., Rankin, K., Fu, H., Weljie, A., Jha, N. and Slupsky, C., 2009. Understanding the Human Salivary Metabolome. *NMR in Biomedicine*, 22(6), pp.577–84.

Tateda, M., Shiga, K., Saijo, S., Sone, M., Hori, T., Yokoyama, J., Matsuura, K., Takasaka, T. and Miyagi, T., 2000. *Streptococcus anginosus* in Head and Neck Squamous Cell Carcinoma: Implication in Carcinogenesis. *International Journal of Molecular Medicine*, 6(6), pp.699–703.

The EUROGAST Study Group, 1993. An International Association Between *Helicobacter pylori* Infection and Gastric Cancer. *The Lancet*, 341(8857), pp.1359–1363.

The Human Microbiome Consortium, 2012a. A Framework for Human Microbiome Research. *Nature*, 486(7402), pp.215–21.

The Human Microbiome Consortium, 2012b. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*, 486(7402), pp.207–14.

Travis, W.D., Brambilla, E., Muller-Hermelink, H.K., Harris, C.C. and (Eds.), 2004. *World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart*.

Turnbaugh, P.J. and Gordon, J.I., 2008. An Invitation to the Marriage of Metagenomics and Metabolomics. *Cell*, 134(5), pp.708–13.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., Egholm, M., Henrissat, B., Heath, A.C., Knight, R. and Gordon, J.I., 2009. A Core Gut Microbiome in Obese and Lean Twins. *Nature*, 457(7228), pp.480–U7.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., Mecklin, J.-P., Järvinen, H., Ristimäki, A., Di-Bernardo, M., East, P., Carvajal-Carmona, L., Houlston, R.S., Tomlinson, I., Palin, K., Ukkonen, E., Karhu, A., Taipale, J. and Aaltonen, L.A., 2009. The Common Colorectal Cancer Predisposition SNP rs6983267 at Chromosome 8q24 Confers Potential to Enhanced Wnt Signaling. *Nature Genetics*, 41(8), pp.885–90.

Ubhi, B.K., Riley, J.H., Shaw, P.A., Lomas, D.A., Tal-Singer, R., MacNee, W., Griffin, J.L. and Connor, S.C., 2012. Metabolic Profiling Detects Biomarkers of Protein Degradation in COPD Patients. *The European Respiratory Journal*, 40(2), pp.345–55.

Uemura, N., Okamoto, S., Yamamoto, S., Matsumura, N., Yamaguchi, S., Yamakido, M., Taniyama, K., Sasaki, N. and Schlemper, R.J., 2001. Helicobacter pylori Infection and the Development of Gastric Cancer. *The New England Journal of Medicine*, 345(11), pp.784–9.

Varki, A., 2007. Glycan-Based Interactions Involving Vertebrate Sialic-Acid-Recognizing Proteins. *Nature*, 446(7139), pp.1023–9.

Varki, A. and Gagneux, P., 2012. Multifarious Roles of Sialic Acids in Immunity. *Annals of the New York Academy of Sciences*, 1253, pp.16–36.

Vartoukian, S.R., Palmer, R.M. and Wade, W.G., 2010. Strategies for Culture of 'Unculturable' Bacteria. *FEMS Microbiology Letters*, 309(1), pp.1–7.

Vaught, J.B., 2006. Blood Collection, Shipment, Processing, and Storage. *Cancer Epidemiology Biomarkers and Prevention*, 15(9), pp.1582–4.

Venkateswaran, K., Vaishampayan, P., Cisneros, J., Pierson, D.L., Rogers, S.O. and Perry, J., 2014. International Space Station Environmental Microbiome - Microbial Inventories of ISS Filter Debris. *Applied Microbiology and Biotechnology*, 98(14), pp.6453–66.

Vimr, E. and Lichtensteiger, C., 2002. To Sialylate, or not to Sialylate?: That is the Question. *Trends in Microbiology*, 10(6), pp.254–7.

Vishnivetskaya, T.A., Layton, A.C., Lau, M.C.Y., Chauhan, A., Cheng, K.R., Meyers, A.J., Murphy, J.R., Rogers, A.W., Saarunya, G.S., Williams, D.E., Pfiffner, S.M., Biggerstaff, J.P., Stackhouse, B.T., Phelps, T.J., Whyte, L., Sayler, G.S. and Onstott, T.C., 2014. Commercial DNA Extraction Kits Impact Observed Microbial Community Composition in Permafrost Samples. *FEMS Microbiology Ecology*, 87(1), pp.217–30.

De Vos, W.M. and de Vos, E.A.J., 2012. Role of the Intestinal Microbiome in Health and Disease: From Correlation to Causation. *Nutrition Reviews*, 70 Suppl 1, pp.S45–56.

Voynow, J.A. and Rubin, B.K., 2009. Mucins, Mucus, and Sputum. *Chest*, 135(2), pp.505–12.

Vuckovic, D., 2012. Current Trends and Challenges in Sample Preparation for Global Metabolomics using Liquid Chromatography-Mass Spectrometry. *Analytical and Bioanalytical Chemistry*, 403(6), pp.1523–48.

Wagner Mackenzie, B., Waite, D.W. and Taylor, M.W., 2015. Evaluating Variation in Human Gut Microbiota Profiles Due to DNA Extraction Method and Inter-Subject Differences. *Frontiers in Microbiology*, 6, p.130.

Warburg, O., 1956. On the Origin of Cancer Cells. *Science*, 123(3191), pp.309–314.

Washio, J., Mayanagi, G. and Takahashi, N., 2010. Challenge to Metabolomics of Oral Biofilm. *Journal of Oral Biosciences*, 52(3), pp.225–232.

Wedzicha, J.A. and Seemungal, T.A.R., 2007. COPD Exacerbations: Defining Their Cause and Prevention. *Lancet*, 370(9589), pp.786–96.

Wei, J., Xie, G., Zhou, Z., Shi, P., Qiu, Y., Zheng, X., Chen, T., Su, M., Zhao, A. and Jia, W., 2011. Salivary Metabolite Signatures of Oral Cancer and Leukoplakia. *International Journal of Cancer*, 129(9), pp.2207–17.

Wei, Q.Y., Cheng, L., Amos, C.I., Wang, L.E., Guo, Z.Z., Hong, W.K. and Spitz, M.R., 2000. Repair of Tobacco Carcinogen-Induced DNA Adducts and Lung Cancer Risk: A Molecular Epidemiologic Study. *Journal of the National Cancer Institute*, 92(21), pp.1764–1772.

Weinberger, M. and Abu-Hasan, M., 2007. Pseudo-Asthma: When Cough, Wheezing and Dyspnea Are Not Asthma. *Pediatrics*, 120(4), pp.855–864.

Weinstock, G.M., 2012. Genomic Approaches to Studying the Human Microbiota. *Nature*, 489(7415), pp.250–6.

Wickström, C. and Svensäter, G., 2008. Salivary Gel-Forming Mucin MUC5B - A Nutrient for Dental Plaque Bacteria. *Oral Microbiology and Immunology*, 23(3), pp.177–82.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E.M., Kyrpides, N., Mavrommatis, K. and Meyer, F., 2012. The M5nr: A Novel Non-Redundant Database Containing Protein Sequences and Annotations from Multiple Sources and Associated Tools. *BMC Bioinformatics*, 13(1), p.141.

Wilkinson, T.M.A., Patel, I.S., Wilks, M., Donaldson, G.C. and Wedzicha, J.A., 2003. Airway Bacterial Load and FEV1 Decline in Patients with Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, 167(8), pp.1090–5.

Willing, B.P., Dicksved, J., Halfvarson, J., Andersson, A.F., Lucio, M., Zheng, Z., Järnerot, G., Tysk, C., Jansson, J.K. and Engstrand, L., 2010. A Pyrosequencing Study in Twins Shows that Gastrointestinal Microbial Profiles Vary with Inflammatory Bowel Disease Phenotypes. *Gastroenterology*, 139(6), pp.1844–1854.e1.

Willner, D., Daly, J., Whiley, D., Grimwood, K., Wainwright, C.E. and Hugenholtz, P., 2012a. Comparison of DNA Extraction Methods for Microbial Community Profiling with an Application to Pediatric Bronchoalveolar Lavage Samples. *PLoS One*, 7(4), p.e34605.

Willner, D., Haynes, M.R., Furlan, M., Hanson, N., Kirby, B., Lim, Y.W., Rainey, P.B., Schmieder, R., Youle, M., Conrad, D. and Rohwer, F., 2012b. Case Studies of the Spatial Heterogeneity of DNA Viruses in the Cystic Fibrosis Lung. *American Journal of Respiratory Cell and Molecular Biology*, 46(2), pp.127–31.

Willner, D., Haynes, M.R., Furlan, M., Schmieder, R., Lim, Y.W., Rainey, P.B., Rohwer, F. and Conrad, D., 2011. Spatial Distribution of Microbial Communities in the Cystic Fibrosis Lung. *ISME Journal*.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-

Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R. and Scalbert, A., 2013. HMDB 3.0: The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(Database issue), pp.D801–7.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J. and Querengesser, L., 2007. HMDB: The Human Metabolome Database. *Nucleic Acids Research*, 35(Database issue), pp.D521–6.

Wittmann, J. and Jäck, H.-M., 2010. Serum microRNAs as Powerful Cancer Biomarkers. *Biochimica et Biophysica Acta*, 1806(2), pp.200–7.

Woo, P.C.-Y., 2003. *Granulicatella adiacens* and *Abiotrophia defectiva* Bacteraemia Characterized by 16S rRNA Gene Sequencing. *Journal of Medical Microbiology*, 52(2), pp.137–140.

Wood, D.E., Eapen, G.A., Ettinger, D.S., Hou, L., Jackman, D., Kazerooni, E., Klippenstein, D., Lackner, R.P., Leard, L., Leung, A.N.C., Massion, P.P., Meyers, B.F., Munden, R.F., Otterson, G.A., Peairs, K., Pipavath, S., Pratt-Pozo, C., Reddy, C., Reid, M.E., Rotter, A.J., Schabath, M.B., Sequist, L. V., Tong, B.C., Travis, W.D., Unger, M. and Yang, S.C., 2012. Lung Cancer Screening. *Journal of the National Comprehensive Cancer Network*, 10(2), pp.240–265.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., Hooper, S.D., Pati, A., Lykidis, A., Spring, S., Anderson, I.J., D'haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J.-F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E.M., Kyrpides, N.C., Klenk, H.-P. and Eisen, J.A., 2009. A Phylogeny-Driven Genomic Encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276), pp.1056–60.

Wu, G.D., Lewis, J.D., Hoffmann, C., Chen, Y.-Y., Knight, R., Bittinger, K., Hwang, J., Chen, J., Berkowsky, R., Nessel, L., Li, H. and Bushman, F.D., 2010. Sampling and Pyrosequencing Methods for

Characterizing Bacterial Communities in the Human Gut using 16S Sequence Tags. *BMC Microbiology*, 10(1), p.206.

Xia, J., Broadhurst, D.I., Wilson, M. and Wishart, D.S., 2013. Translational Biomarker Discovery in Clinical Metabolomics: An Introductory Tutorial. *Metabolomics*, 9(2), pp.280–299.

Xia, J., Mandal, R., Sinelnikov, I. V, Broadhurst, D. and Wishart, D.S., 2012. MetaboAnalyst 2.0: A Comprehensive Server for Metabolomic Data Analysis. *Nucleic Acids Research*, 40(Web Server Issue), pp.W127–33.

Xie, Y., Todd, N.W., Liu, Z., Zhan, M., Fang, H., Peng, H., Alattar, M., Deepak, J., Stass, S.A. and Jiang, F., 2010. Altered miRNA Expression in Sputum for Diagnosis of Non-Small Cell Lung Cancer. *Lung Cancer*, 67(2), pp.170–6.

Xiong, W., Wang, L. and Yu, F., 2014. Regulation of Cellular Iron Metabolism and its Implications in Lung Cancer Progression. *Medical Oncology*, 31(7), p.28.

Xu, J., 2006. Microbial Ecology in the Age of Genomics and Metagenomics: Concepts, Tools, and Recent Advances. *Molecular Ecology*, 15(7), pp.1713–31.

Yadav, A.P., Chaturvedi, S., Mishra, K.P., Pal, S., Ganju, L. and Singh, S.B., 2014. Evidence for Altered Metabolic Pathways During Environmental Stress: 1H-NMR Spectroscopy Based Metabolomics and Clinical Studies on Subjects of Sea-Voyage and Antarctic-Stay. *Physiology and Behaviour*, 135, pp.81–90.

Yang, F., Zeng, X., Ning, K., Liu, K.-L., Lo, C.-C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., Chain, P.S., Xie, G., Ling, J. and Xu, J., 2012. Saliva Microbiomes Distinguish Caries-Active from Healthy Human Populations. *The ISME Journal*, 6(1), pp.1–10.

Zawadzki, M.A., Ihrig, A.M., Grider, D.A., Jessup, T.D. and Williams, D.L., 2005. *Reduced Ignition Propensity Smoking Article (US Patent 6837248 B2)*. US Patent 6837248 B2.

Van der Zee, A., Peeters, M., de Jong, C., Verbakel, H., Crielaard, J.W., Claas, E.C.J. and Templeton, K.E., 2002. Qiagen DNA Extraction Kits for Sample Preparation for *Legionella* PCR are not Suitable for Diagnostic Purposes. *Journal of Clinical Microbiology*, 40(3), pp.1126–1126.

Zhang, A., Sun, H., Wang, P., Han, Y. and Wang, X., 2012. Modern Analytical Techniques in Metabolomics Analysis. *The Analyst*, 137(2), pp.293–300.

Zhao, J., Li, J., Schloss, P.D., Kalikin, L.M., Raymond, T.A., Petrosino, J.F., Young, V.B. and LiPuma, J.J., 2011. Effect of Sample Storage Conditions on Culture-Independent Bacterial Community Measures in Cystic Fibrosis Sputum Specimens. *Journal of Clinical Microbiology*, 49(10), pp.3717–3718.

APPENDIX

Due to their size, Supplementary Tables have been included as digital files on a disc attached to the back cover of this thesis. The Supplementary Table legends have been included here for reference.

Supplementary Table 2.1 | Microbiomics Full Participant Information

Full participant information for ten patients who donated samples as part of the lung cancer microbiome portion of this work. Full participant information includes individual MG-RAST metagenome ID, age, sex, drug and medical history, smoking history, presence of infection, FEV1 % of predicted, and lung cancer state and staging.

Supplementary Table 2.2 | Metagenomic Sequence Read Analysis

Average read statistics for pre- and post-quality control (QC), for each group, alongside one-way ANOVA P values. Analysis shows no significant differences in all bar one, identified rRNA features, suggesting that the sequencing approach, and subsequent analysis using the MG-RAST pipeline, used in this study has not introduced any discernible bias between the two groups.

Supplementary Table 2.3 | Metabolomics Full Participant Information

Full individualised patient clinical information, including drug and medical history for all clinical patients, alongside relevant clinical diagnosis and histology, and information for participants from Swansea University.

Supplementary Table 3.1 | COPD Metagenomics Full Participant Information

Full participant information for Control participants and COPD patients, showing age, gender, and smoking history. Additional clinical information for COPD patients includes drug history, medical history, FEV_1 % of predicted, and whether the patient had an infection at the time of giving a sample. nc = not collected.

Supplementary Table 3.2 | Metagenomic Sequence Read Analysis

Average read statistics pre and post quality control (QC), after merging of paired-end reads, alongside corresponding one-way ANOVA *P* values. Analysis shows no significant differences in all but one read characteristic, average read length both pre and post QC, suggesting that the HiSeq 2500 sequencing approach and MG-RAST analysis pipeline introduced no discernible bias between the two participant groups.

Supplementary Table 4.1 | Illumina Adaptors Used in 16S rRNA Amplicon Sequencing

To allow for multiplexing of all 70 samples in one Illumina MiSeq 2 x 300 bp run, unique barcode sequences were added to the first stage PCR products after amplification of the V3 to V4 region of the 16S rRNA gene. Illumina Nextera XT adaptors were used, and the specific Index 1 (i7) and Index 2 (i5) for each sample is given here.

Supplementary Table 4.2 | Seasonal Variation Full Participant Information

Individual participant information, for each of the seven sampling periods, is given here, alongside additional information collected during the October 2013 sampling period. All data is link anonymised. The individual participant information for those participants whose samples underwent amplicon sequencing of the 16S rRNA gene is also given.

Supplementary Table 4.3 | Sequence Statistics for 16S rRNA Amplicon Sequencing

Sequence statistics for those ten participants whose samples underwent amplicon sequencing of the 16S rRNA gene, grouped by participant and sampling month. All features, except sequence length, GC content, and identified rRNA features in participant groupings, show no significant differences. All values given as means in standard format, except for sequence length and GC content statistics, with standard deviations.

Chapter 5 | Supplementary Information

Supplementary Table 5.1 | Illumina Adaptors Used in 16S rRNA Amplicon Sequencing

To allow for multiplexing of all 90 samples in one Illumina MiSeq 2 x 300 bp run, unique barcode sequences were added to the first stage PCR products after amplification of the V3 to V4 region of the 16S rRNA gene. Illumina Nextera XT adaptors were used, and the specific Index 1 (i7) and Index 2 (i5) for each sample is given here.

Supplementary Table 5.2 | Full Participant Physiological Information

Full participant physiological information, namely weight, body mass index, and body fat percentage is given for each sampling month, which corresponds with when stool, saliva, and blood plasma samples were taken.

Supplementary Table 5.3 | Sequence Statistics for Saliva 16S rRNA Amplicon Sequencing

Sequence statistics for the 45 saliva samples which underwent amplicon sequencing of the 16S rRNA gene, grouped by participant and sampling month. All features, except sequence length, and GC content in participant groupings, show no significant differences. All values given as means in standard format, except for sequence length and GC content statistics, with standard deviations.

Supplementary Table 5.4 | Sequence Statistics for Stool 16S rRNA Amplicon Sequencing

Sequence statistics for the 45 stool samples which underwent amplicon sequencing of the 16S rRNA gene, grouped by participant and sampling month. All features, except processed rRNA features and aligned rRNA features showed some degree of significant difference between participant groupings, but all showed no significant differences in regards to sampling month groupings. All values given as means in standard format, except for sequence length and GC content statistics, with standard deviations.